The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Label Distribution Learning by Optimal Transport*

Peng Zhao, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China
{zhaop, zhouzh}@lamda.nju.edu.cn

Abstract

Label distribution learning (LDL) is a novel learning paradigm to deal with some real-world applications, especially when we care more about the relative importance of different labels in description of an instance. Although some approaches have been proposed to learn the label distribution, they could not explicitly learn and leverage the label correlation, which plays an importance role in LDL. In this paper, we propose an approach to learn the label distribution and exploit label correlations simultaneously based on the *Optimal Transport* (OT) theory. The problem is solved by alternatively learning the transportation (hypothesis) and ground metric (label correlations). Besides, we provide perhaps the first data-dependent risk bound analysis for label distribution learning by Sinkhorn distance, a commonly-used relaxation for OT distance. Experimental results on real-world datasets comparing with several state-of-the-art methods validate the effectiveness of our approach.

Introduction

In traditional machine learning paradigm, an instance is in general associated to a single label. Single-label learning (SLL) is established to deal with this case. However, as is often the case, multiple labels might be linked with the same instance simultaneously. Taking the image classification as an example, an image can be complicated and have multiple semantic meanings. Consequently, it can be tagged with several different categories/tags. To handle such kind of tasks, multi-label learning (MLL) paradigm is proposed and has achieved a lot of success (Zhang and Zhou 2007; 2014; Zhou et al. 2012).

SLL and MLL only answer the question which label/labels should the instance belong to, but not the relative importance of different labels in description of the instance. However, various labels may be associated to an instance with different degrees in many real-world applications. Thus, it is more reasonable to use a soft label description rather a hard one. Label distribution learning (LDL) (Geng 2016) is a new learning paradigm to describe supervision as a histogram or probability distribution.

*This research was supported by NSFC (61333014) and the 973 Program (2014CB340501).

LDL has been successfully applied in many real-world scenarios in recent years, such as facial age estimation (Geng, Yin, and Zhou 2013; Gao et al. 2017), head pose estimation (Kong and Mbouna 2015; Xu and Zhou 2017), emotion recognition (Zhou, Xue, and Geng 2015) and text mining (Zhou et al. 2016). Although LDL has been wildly applied in different scenarios and achieves a great success, most of previous LDL approaches cannot automatically exploit the label correlations to boost the learning performance. An exception is a recent work (Zhou et al. 2016) dealing with emotion analysis from texts. It exploits the relationship between different emotions by adding a specific-designed regularization term based on prior domain knowledge. However, in many other applications, there is usually no such prior knowledge or additional structure information on label correlations. Moreover, the specific-designed regularizer needs to be redesigned for new applications.

In this work, we aim to simultaneously learn the label distribution and explore label correlations. Thus, we proposed an algorithm called **LALOT**, short for **LA**bel Distribution **L**earning by **O**ptimal **T**ransport. This method is based on *Optimal Transport* (OT) theory (Villani 2008). On one hand, we adopt optimal transport distance to measure the quality of prediction. OT distance provides a more meaningful distance in LDL tasks, because it could capture the geometric information of the underlying label space. On the other hand, we cast the label correlations exploration as a ground metric learning problem. The ground metric is also called cost matrix, playing an important role in the performance of optimal transport. Previous work mainly assumed that there exists some prior knowledge as cost matrix (Frogner et al. 2015; Rolet, Cuturi, and Peyré 2016). However, this may not hold in many real-world applications, since there may not be a direct or simple semantic information among labels. LALOT could automatically learn the ground metric, i.e., the label correlations, and improve the label distribution learning performance. It is also noteworthy to mention that most previous work on LDL do not have any theoretical analysis, and we provide perhaps the first risk bound analysis for label distribution learning in this work.

The main contributions of this paper are summarized in following three points,

1) We cast LDL as solving an optimal transport problem, and handle label correlations exploration by ground met-

ric learning. The optimal transport and ground metric are jointly learned by alternative optimization.

2) We do not require a prior structure knowledge on labels but directly learn the ground metric by kernel biased regularization. Our approach avoids a costly projection to metric space, only projection to semi-definite cone is necessary, whose projection cost is much cheaper.

3) We provide, to the best of our knowledge, the first data-dependent risk bound analysis for label distribution learning. Besides, this might also be the first risk bound analysis for Sinkhorn distance, a commonly used entropic regularized approximation for OT.

In the following, we start with a brief review of related work. Then, the proposed approach LALOT will be introduced. Next, both theoretical results and empirical effectiveness have been examined. Finally, we conclude the paper.

Related Work

Label Distribution Learning (LDL) (Geng 2016) is a novel machine learning framework. It models multiple labels as a label distribution indicating the relative importance of each label involved in the description of an instance.

There are several algorithms designed for LDL. In general, they can be grouped into three categories: problem transformation, algorithm adaptation and specialized algorithms. Problem transformation method aims to change the training examples into weighted single-label instances by sampling in order to transform LDL problem as SLL or MLL learning. After the sampling, SVM and Naive Bayes can be applied to perform the binary classification, developing two representative algorithms called PT-SVM and PT-Bayes respectively (Geng 2016). Algorithm adaptation methods extend traditional MLL algorithms to deal with label distributions, they adapt the hard-threshold labels to the soft ones by some specific mechanisms, such as AA-Bayes and AA-BP (Geng 2016). Besides, there are some specialized algorithms to directly match the LDL problems, a representative one is IIS-LLD (Geng, Yin, and Zhou 2013).

There are few papers considering label distribution learning from the theoretical aspect. In this paper, we develop a first data-dependent risk analysis for label distribution learning based on Rademacher complexity.

Optimal Transport (OT) (Villani 2008) is originally developed to measure the difference between two probability distributions based on some given ground metric. The distance defined by OT is also called Wasserstein distance or Earth Mover's distance for some special cases. Recently, OT has drawn great attentions in computer vision and image processing fields, such as image retrieval (Rubner, Tomasi, and Guibas 2000), barycenters computation (Cuturi and Doucet 2014). Also, it is wildly applied in machine learning and related fields, for instance NMF (Qian et al. 2016), clustering (Ye et al. 2017), domain adaptation (Courty et al. 2016) and multi-label learning (Frogner et al. 2015).

Most previous work require a similarity structure in the output space as a prior side information to define the optimal transport distance. Only a few work (Cuturi and Avis 2014; Rolet, Cuturi, and Peyré 2016) try to learn the ground metric. However, they share two common drawbacks. The first concern is that they have to bare a high computational cost due to a projection back to metric space. The other issue is that they still need some additional information on similar or dissimilar samples, although they do not explicitly assume a similarity structure.

Different from previous work, we do not require prior knowledge on the label structures or additional information on samples, but jointly learn the transportation (hypothesis) and ground metric (label correlations). Moreover, by the mechanism of kernel biased regularization, the projection cost of our approach is much cheaper, since only a projection to semi-definite cone is necessary.

Preliminary

Notations

For two matrices $X, Y \in \mathbb{R}^{m\times n}$, $\langle X, Y\rangle \stackrel{\text{def}}{=} \mathrm{tr}(X^{\mathrm{T}}Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij}$ is the Frobenius dot-product, $X \oslash Y$ denotes the element-wise quotient between X and Y. Besides, we denote the simplex as $\Sigma_d := \{x \in \mathbb{R}_+^d : x^{\mathrm{T}}\mathbf{1}_d = 1\}$, where $\mathbf{1}_d$ is the d dimensional vector of ones.

Optimal Transport Distance

It is of great importance to define a proper distance between two probability distribution, which is key to many machine learning problems. Traditionally, plenty of measurements are introduced including Hellinger, total variation and Kullback-Leibler divergences. However, these measurements fail when the probability space has some geometrical structures. The optimal transport distances are usually more powerful in such situations, since it could take the pairwise cost into consideration when measuring the distance.

Definition 1. *(Transport Polytope) For two probability vectors r and c in the simplex Σ_d, we write $U(r, c)$ for the transport polytope of r and c, namely the polyhedral set of $d \times d$ matrices,*

$$U(r, c) := \{P \in \mathbb{R}_+^{d\times d} | P\mathbf{1}_d = r, P^{\mathrm{T}}\mathbf{1}_d = c\}. \quad (1)$$

Definition 2. *(Optimal Transport) Given a $d \times d$ cost matrix M, the total cost of mapping from r to c using a transport matrix (or coupling probability) P can be quantified as $\langle P, M\rangle$. The optimal transport (OT) problem is defined as,*

$$d_M(r, c) := \min_{P\in U(r,c)} \langle P, M\rangle. \quad (2)$$

Theorem 1. *(Optimal Transport Distance (Villani 2008)) d_M defined in (2) is a distance on Σ_d whenever M is ~~a metric matrix.~~*

Remark 1. OT distance is also known as the Earth Mover's distance (Rubner, Tomasi, and Guibas 2000). It is very similar to Wasserstein distance (Bogachev and Kolesnikov 2012), but they have some subtle differences. The cost matrix of Wasserstein distance is defined as $M_{ij} = d_{\mathcal{K}}^p(i, j)$, where $d_{\mathcal{K}}$ is a metric and p is an integer greater than 1. As a matter of fact, its cost matrix may not be a metric.

The Proposed Approach

Problem Statements

We consider the problem of learning a mapping from the feature space $\mathcal{X}$ to the label space $\mathcal{Y}$. Let the training set be $S = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_m, \mathbf{y}_m)\}$, sampled from an underlying probability distribution $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. We denote $X = [\mathbf{x}_1, \cdots, \mathbf{x}_m]^{\mathrm{T}} \in \mathbb{R}^{m\times d}$ and $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_m]^{\mathrm{T}}$, where $\mathbf{y}_i \in \Sigma_L$ and L is the total number of labels. The goal is to learn label distribution and explore label correlations simultaneously.

Label Correlations Exploration

Non-linear Transformation First, we introduce a non-linear transformation $\phi(\cdot)$, and denote the Euclidean distance after the transformation as

$$\mathcal{D}_\phi(\mathbf{u}, \mathbf{v}) = \|\phi(\mathbf{u}) - \phi(\mathbf{v})\|_2. \tag{3}$$

$\mathcal{D}_\phi$ satisfies all properties of a well-defined pseudo-metric in the original input space (Kedem et al. 2012).

Remark 2. Technically, the non-linear transformation defined in (3) is not a strictly metric, but a pseudo-metric (also named as semi-metric). Thus, OT defined accordingly is no longer a strict distance. However, it preserves sub-additivity property, which plays a key role in measuring difference between prediction and groundtruth. Meanwhile, it is sufficient to make it a strict distance by multiplying d_M by $\mathbf{1}_{r\neq c}$.

Theorem 2. *For a pseudo-metric M and probability distributions $r, c \in \Sigma_d$, the function $(r, c) \to \mathbf{1}_{r\neq c} d_M(r, c)$ satisfies all four distance axioms, i.e., non-negativity, symmetry, definiteness and sub-additivity (triangle inequality).*

The proof follows (Cuturi and Avis 2014), and the key idea is to exploit the sub-additivity of cost matrix, and detailed proof will be presented in longer version.

Kernel Biased Regularization Instead of directly learning the ground metric, we adopt kernel biased term as regularizer. Specifically, we learn a kernel K defined by the non-linear transformation $\phi(\cdot)$, namely,

$$K_{ij} = K(Y_{:,i}, Y_{:,j}) = \phi(Y_{:,i})^{\mathrm{T}} \phi(Y_{:,j}), \tag{4}$$

where $Y_{:,i}$ denotes the i-th column of label matrix Y.

The Formulation

Now, we propose to conduct the label distribution learning and label correlations exploration simultaneously based on optimal transport. We adopt OT mainly because OT distance can be used to capture the geometry of a space. To do so, we adopt OT distance as the loss between prediction and groundtruth, and then incorporate the ground metric learning by kernel biased regularization,

$$\begin{aligned} \min_{K, h\in\mathcal{H}} \quad & \sum_{i=1}^{m} \langle P_i, M\rangle + \frac{C}{2}\|K - K_0\|_F^2 \\ \text{s.t.} \quad & P_i \in U(h(\mathbf{x}_i), \mathbf{y}_i) \\ & K \in \mathcal{S}_+, \end{aligned} \tag{5}$$

where $C > 0$ is a trade-off parameter, $\mathcal{S}_+$ denotes the set of positive semi-definite matrices, $\mathcal{H}$ is hypothesis set and $U(h(\mathbf{x}_i), \mathbf{y}_i)$ is defined in Definition 1. Besides, the ground metric M is computed as $M_{ij} = \mathcal{D}_\phi^2(Y_{:,i}, Y_{:,j})$, and the kernel K is defined as (4). Thus, the relation between M and K can be derived as

$$M_{ij} = K_{ii} - 2K_{ij} + K_{jj}. \tag{6}$$

Remark 3. The non-linear mapping preserves (pseudo-) metric properties, and therefore it only needs a projection to positive semi-definite matrix cone when learning the kernel. Thus, we can avoid the projection to ~~metric space~~ which is very complicated and costly.

Remark 4. It seems that we could learn the cost matrix and hypothesis jointly by $(M^\star, h^\star) = \arg\min_{M,h\in\mathcal{H}} \langle P, M\rangle$. In fact, this will be an ill-posed optimization problem, since it has a trivial solution as $M = \mathbf{0}$ along with arbitrary h. Consequently, the kernel biased regularizer in (5) is necessary for ground metric learning. The idea is similar to the biased regularization in hypothesis transfer learning (Kuzborskij and Orabona 2017). In this paper, we adopt the label distribution covariance matrix as the initialization, namely, $K_0 = Y^{\mathrm{T}}Y$. If there are auxiliary convincing label correlations, it can be transformed into the initial kernel or directly used as ground metric. More discussions on initialization strategy will be appeared in longer version.

Optimization

We adopt the alternative optimization to solve problem (5),

(i) fix K to update h: learning the target mapping;

(ii) fix h to update K: learning the ground metric.

Learning the Target Mapping When updating h with a fixed K, the sub-problem can be written as follows,

$$\begin{aligned} \min_{h\in\mathcal{H}} \quad & \sum_{i=1}^{m} \langle P_i, M\rangle \\ \text{s.t.} \quad & P_i \in U(h(\mathbf{x}_i), \mathbf{y}_i). \end{aligned} \tag{7}$$

The sub-problem can be solved by gradient descent, however, it is a challenge to directly compute the gradient w.r.t. prediction $h(\mathbf{x}_i)$, especially when it is in the constraints.

In this paper, similar to (Frogner et al. 2015), we use primal-dual approach to compute its gradient by solving the dual LP problem, and adopt Sinkhorn's relaxation (Cuturi 2013) as the entropic regularization to smooth the transport objective and speed up the computation of original OT.

For a given training sample $(\mathbf{x}, \mathbf{y})$, the dual LP of (2) is

$${}^d d_M(h(\mathbf{x}), \mathbf{y}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}\in C_M} \boldsymbol{\alpha}^{\mathrm{T}} h(\mathbf{x}) + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{y},$$

where $C_M = \{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^L : \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j \leq M_{i,j}\}$.

From (Bertsimas and Tsitsiklis 1997), we know that the dual optimal $\boldsymbol{\alpha}$ is, in fact, a subgradient of the loss of training sample $(\mathbf{x}, \mathbf{y})$ with respect to its first argument $h(\mathbf{x})$. However, it is costly to directly compute the exact loss. In the seminal paper (Cuturi 2013), Cuturi introduced an entropic regularization as an efficient approximation of original problem named as the *Sinkhorn distance*.

Definition 3. *(Sinkhorn Distance) Given a $d \times d$ cost matrix M, and marginal distributions $r, c \in \Sigma_d$. The Sinkhorn distance is defined as,*

$$d_M^\lambda(r,c) := \langle P^\lambda, M\rangle, \tag{8}$$

$$P^\lambda = \underset{P\in U(r,c)}{\arg\min} \langle P, M\rangle - \frac{1}{\lambda}H(P), \tag{9}$$

where $H(P) = -\sum_{i=1}^d \sum_{j=1}^d p_{ij}\log p_{ij}$ is the entropy of P, and $\lambda > 0$ is entropic regularization coefficient.

The advantages of entropic regularization are multi-faceted, and the most important one is that the entropic regularization makes the objective function a strictly convex problem that can be solved through Sinkhorn's matrix scaling algorithm, at a speed that is several orders of magnitude faster than that of transport solvers (Cuturi 2013).

Then, based on the Sinkhorn's theorem, we could conclude that the transportation matrix can be written in the form of $P^\star = \mathtt{diag}(\mathbf{u})\mathbf{K}\mathtt{diag}(\mathbf{v})$, where $\mathbf{K} := e^{-\lambda M}$ is the element-wise exponential of λM. Besides, $\mathbf{u} = e^{-1/2+\lambda\boldsymbol{\alpha}}$ and $\mathbf{v} = e^{-1/2+\lambda\boldsymbol{\beta}}$.

Thus, we adopt the well-known Sinkhorn-Knopp algorithm which is also used in (Cuturi 2013; Cuturi and Doucet 2014) to update the target mapping $h(\mathbf{x})$ given the ground metric, i.e., with a fixed K. In this paper, we adopt linear logistic regression as base classifier $h(\mathbf{x})$, defined as

$$\hat{\mathbf{y}}_k = h(\mathbf{x}_k) = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_L]^{\mathrm{T}}, \hat{y}_i = \frac{\exp\left(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}_k\right)}{\sum_{i=1}^L \exp\left(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}_k\right)},$$

where $\hat{y}_i$ is the i-th dimensional value of prediction.

To simplify the notations, we denote the object function as $F(W) = \sum_{k=1}^m \ell(h(\mathbf{x}_k), \mathbf{y}_k)$, where $W = [\mathbf{w}_1; \cdots; \mathbf{w}_L]^{\mathrm{T}} \in \mathbb{R}^{L\times d}$. Then the gradient w.r.t. W could be calculated via chain rule.

The detailed procedure is summarized in Algorithm 1, then gradient descent could be adopted to update classifier. Obviously, for datasets with large amount of instances, it could be easily extended to SGD version for acceleration.

Learning the Ground Metric When updating K with a fixed h, the sub-problem can be written as follows,

$$\begin{aligned} \min_K \quad & \langle P, M\rangle + \frac{C}{2}\|K - K_0\|_F^2 \\ \text{s.t.} \quad & K \in \mathcal{S}_+ \\ & M_{ij} = K_{ii} + K_{jj} - 2K_{ij}, \end{aligned} \tag{10}$$

where $P = \sum_{i=1}^m P_i$.

This sub-problem can be solved by projected gradient descent. To simplify the notations, let $G(K)$ be the objective function, i.e.,

$$G(K) = \frac{C}{2}\|K - K_0\|_F^2 + g(K),$$

where $g(K) = \langle P, M\rangle$. Since $\nabla G(K) = C(K - K_0) + \nabla g(K)$, we could turn to compute $\nabla g(K)$, and let the full gradient be zero, obtaining the close-form solution as,

Algorithm 1 Learning the Mapping

Input: Ground metric M, current mapping $h(\mathbf{x})$, training set $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)\}$, and $\lambda > 0$;
Output: Gradient of objective function with respect to the target mapping $h(x)$.

```
Initialize u ← 1, K = e^{-λM}, ∇ ← 0;
for i = 1 to m do
    u_i ← 1;
    while u_i has not converged do
        u_i ← h(x_i) ⊘ (K(y_i ⊘ K^T u_i));
    end while
    ∇_i^H ← (log u_i)/λ − (log u_i^T 1)/(λL) · 1;
    Compute ∇_i^W according to the chain rule;
    ∇ ← ∇ + [∇_i^H]^T ∇_i^W;
end for
```

$$(\hat{K} - K_0)_{ij} = \begin{cases} 2P_{ij} & \text{, when } i \neq j \\ -\sum\limits_{k\neq i}^{L}(P_{ik} + P_{ki}) & \text{, when } i = j \end{cases} \tag{11}$$

Then, we project $\hat{K}$ back to positive semi-definite cone as,

$$K = \mathbf{Proj}(\hat{K}) = U\max(\sigma, 0)U^{\mathrm{T}},$$

where **Proj** is a projection operator, U and σ correspond to the eigenvectors and eigenvalues of $\hat{K}$.

Theoretical Results

In this part, we provide a risk bound analysis for learning the mapping during the alternative optimization. To the best of our knowledge, this might be the first risk analysis for Sinkhorn distance, a relaxation for optimal transport distance in common use. Also, this might be the first data-dependent risk analysis for label distribution learning.

Due to the page limits, only some proof sketches for main theorem are provided, details of the omitted proofs will be presented in longer version.

To simplify the presentation, we introduce the *Sinkhorn loss* as:

$$\ell(h(\mathbf{x}), \mathbf{y}) := d_M^\lambda(h(\mathbf{x}), \mathbf{y}) = \langle P^\lambda, M\rangle, \tag{12}$$

where P^λ is obtained by Sinkhorn iteration defined in (9).

Based on Sinkhorn loss defined in(12), we introduce notations of corresponding risk and empirical risk, respectively.

$$R(h) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\ell(h(\mathbf{x}), \mathbf{y}), \quad \hat{R}(h) = \sum_{i=1}^m \ell(h(\mathbf{x}_i), \mathbf{y}_i).$$

In the following, we will utilize the notion of Rademacher complexity (Bartlett and Mendelson 2002) to measure the hypothesis complexity and use it to bound the excess risk.

Definition 4. *(Rademacher Complexity (Bartlett and Mendelson 2002)) Let $\mathcal{G}$ be a family of functions and a fixed sample of size m as $S = (\mathbf{z}_1, \cdots, \mathbf{z}_m)$. Then, the empirical*

Rademacher complexity of $\mathcal{G}$ with respect to the sample S is defined as:

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i g(\mathbf{z}_i)\right].$$

Besides, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$:

$$\mathfrak{R}_m(\mathcal{G}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\hat{\mathfrak{R}}_S(\mathcal{G})]. \tag{13}$$

Then, we could establish a generalization bound based on Rademacher complexity defined in (13).

Theorem 3. *(Mohri, Rostamizadeh, and Talwalkar 2012) Let $\mathcal{L}$ be the family of loss function associated to $\mathcal{H}$, i.e., $\mathcal{L} = \{\ell(h(\mathbf{x},\mathbf{y}), h \in \mathcal{H}\}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$:*

$$R(h) \le \hat{R}(h) + 2\mathfrak{R}_m(\mathcal{L}) + \|M\|_\infty\sqrt{\frac{\log(1/\delta)}{2m}} \tag{14}$$

where $\mathfrak{R}_m(\mathcal{L})$ is Rademacher complexity of loss function class $\mathcal{L}$ associated to $\mathcal{H}$, and $\|M\|_\infty = \max_{ij} M_{ij}$.

We provide some useful properties of Sinkhorn distance.

Lemma 1. *For any double stochastic matrix $S \in \mathbb{R}_+^{d\times d}$, its entropy $H(S)$ satisfies $H(S) \le 2\log d$.*

Lemma 2. *For two probability distributions $r, c \in \Sigma_d$, Sinkhorn distance $d_M^\lambda(r,c)$ and optimal transport distance $d_M(r,c)$ satisfy the following relationship,*

$$d_M(r,c) \le d_M^\lambda(r,c) \le d_M(r,c) + \frac{2}{\lambda}\log d. \tag{15}$$

In order to establish the relationship between Rademacher complexity of Sinkhorn distance loss and function space, we need introduce another loss definition based on original optimal transport distance as

$$\ell_{OT}(h(\mathbf{x}),\mathbf{y}) := d_M(h(\mathbf{x}),\mathbf{y}) = \langle P, M\rangle. \tag{16}$$

Then, based on Lemma 2, we know that

$$\ell_{OT}(h(\mathbf{x}),\mathbf{y}) \le \ell(h(\mathbf{x}),\mathbf{y}) \le \ell_{OT}(h(\mathbf{x}),\mathbf{y}) + \frac{\log L}{\lambda}.$$

holds for any instance $(\mathbf{x},\mathbf{y})$. Now, we can relate the Rademacher complexity associated with these two losses as stated in Theorem 4.

Theorem 4. *Let $\mathcal{L}$ and $\mathcal{L}_{OT}$ correspond the family of loss functions ℓ and ℓ_{OT} associated to function space $\mathcal{H}$. Then the Rademacher complexities of $\mathcal{L}$ and $\mathcal{L}_{OT}$ satisfy,*

$$\mathfrak{R}_m(\mathcal{L}) \le \mathfrak{R}_m(\mathcal{L}_{OT}) + \frac{\log L}{2\lambda}. \tag{17}$$

Now, we can provide the risk bound for ERM based on Sinkhorn loss.

Theorem 5. *Let $\mathcal{H}$ be the family of hypothesis set, and denote the hypothesis returned by LALOT in Algorithm 1 as $\hat{h}$. Then, for any $\delta > 0$, with probability at least $1-\delta$,*

$$R(\hat{h}) \le \inf_{h\in\mathcal{H}} R(h) + \frac{2\log L}{\lambda} + \|M\|_\infty\left(16L\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{2\log\frac{1}{\delta}}{m}}\right).$$

where $\mathfrak{R}_m(\mathcal{H})$ is Rademacher complexity of hypothesis class $\mathcal{H}$, and $\|M\|_\infty = \max_{ij} M_{ij}$.

Table 1: Statistics of 15 real-world datasets

Index	Datasets	#instance	#dim	#label
1	JAFFE-5	181	18	5
2	JAFFE-6	213	18	6
3	Emotions	593	72	6
4	Image	2000	135	5
5	Yeast-alpha	2,465	24	18
6	Yeast-cdc	2,465	24	15
7	Yeast-cold	2,465	24	4
8	Yeast-diau	2,465	24	7
9	Yeast-dtt	2,465	24	4
10	Yeast-elu	2,465	24	14
11	Yeast-heat	2,465	24	6
12	Yeast-spo	2,465	24	6
13	Yeast-spo5	2,465	24	3
14	Yeast-spoem	2,465	24	2
15	Human Gene	30,542	36	68

Proof Sketch. To prove the risk bound, we turn to establish a uniform generalization bound similar to Theorem 3. Then we use Lemma 2 to connect the relationship between Sinkhorn loss and optimal transport loss.

The main technique is to utilize the concentration of measure, with Rademacher vector contraction inequality applied on a Lipschitz loss, then we can link the Rademacher complexity of loss family and hypothesis space. □

Remark 5. From Eq. (2), (8) and (9), we can see that the Sinkhorn distance coincides with optimal transport distance as $\lambda \to \infty$. Besides, from Theorem 5, we can see that there is a constant error in risk bound but will be reduced to the standard convergence rate $O(1/\sqrt{m})$ as $\lambda \to \infty$, which also coincides with the risk bound for optimal transport distance. This reflects the trade-off between computational efficiency and approximation accuracy to some extents.

Remark 6. From Theorem 5, we can see that the risk bound gets worse with the entropic regularization, namely, when λ is small. This might somehow contradict with the empirical phenomenon reporting that Sinkhorn distance usually performs better than original OT even with a small λ (Cuturi 2013; Cuturi and Doucet 2014). The reason could lie in the bound is not tight, since we have no more advanced tools but *Talagrand's Comparison Inequalities* (Koltchinskii 2011) or *Rademacher Vector Contraction Inequality* (Maurer 2016) to analyze the risk bounds more meticulously. One conjecture is that the entropic regularization could strengthen the convexity of objective function, which may speed up the convergence rate. However, there is no such related work as far as we know. To the best of our knowledge, Theorem 5 provided in this paper is the only result of risk bound analysis for Sinkhorn distance.

Experiments

In this part, we evaluate the effectiveness of LALOT in following two aspects,

(i) label distribution learning: we will examine the effectiveness of LALOT in boosting performance of LDL;

Table 2: Experimental results on LDL datasets. Each row corresponds to a data set. On each dataset, 10 test runs were conducted and the average performance as well as standard deviation are presented, - indicates numerical limits or errors. Besides, • (○) indicates that LALOT is significantly better (worse) than the compared method (paired t-tests at 95% significance level).

(a) Performance Measure: `Chebyshev` ↓

Dataset	IIS-LLD	PT-Bayes	PT-SVM	AA-BP	AA-KNN	LALOT
Image	.6749 ± .0065	.6682 ± .0103	.7305 ± .0465 •	.5863 ± .0370 ○	**.5130 ± .0074** ○	.6850 ± .0073
JAFFE-5	.1396 ± .0046	.4251 ± .0270 •	.1430 ± .0053 •	.1490 ± .0073 •	.1444 ± .0059 •	**.1394 ± .0044**
JAFFE-6	.1207 ± .0060	.3674 ± .0296 •	.1220 ± .0057 •	.1265 ± .0085 •	.1321 ± .0042 •	**.1206 ± .0049**
Emotions	.4429 ± .0138	.6659 ± .0312 •	.5526 ± .0565 •	**.4087 ± .0140** ○	.3989 ± .0155 ○	.4476 ± .0153
Yeast-alpha	.0201 ± .0002 •	.1093 ± .0085 •	.0139 ± .0003	.0358 ± .0022 •	.0147 ± .0002 •	**.0136 ± .0002**
Yeast-cdc	.0232 ± .0005 •	.1211 ± .0080 •	.0170 ± .0005	.0370 ± .0019 •	.0176 ± .0003 •	**.0168 ± .0003**
Yeast-cold	.0618 ± .0007 •	.2060 ± .0155 •	.0565 ± .0033 •	.0574 ± .0024 •	.0553 ± .0009 •	**.0541 ± .0009**
Yeast-diau	.0452 ± .0007 •	.1793 ± .0126 •	.0439 ± .0026 •	.0471 ± .0015 •	**.0395 ± .0006** ○	.0418 ± .0007
Yeast-dtt	.0491 ± .0010 •	.2019 ± .0188 •	.0380 ± .0016 •	.0443 ± .0022 •	.0391 ± .0007 •	**.0370 ± .0006**
Yeast-elu	.0239 ± .0004 •	.1254 ± .0076 •	.0171 ± .0004 •	.0363 ± .0015 •	.0177 ± .0002 •	**.0167 ± .0002**
Yeast-heat	.0526 ± .0006 •	.1942 ± .0086 •	.0441 ± .0009 •	.0520 ± .0013 •	.0453 ± .0003 •	**.0435 ± .0005**
Yeast-spo	.0653 ± .0010 •	.1855 ± .0112 •	.0625 ± .0019 •	.0664 ± .0035 •	.0636 ± .0008 •	**.0603 ± .0012**
Yeast-spo5	.0958 ± .0022 •	.2209 ± .0159 •	.0923 ± .0025 •	.0927 ± .0021 •	.0956 ± .0014 •	**.0908 ± .0015**
Yeast-spoem	.0930 ± .0019 •	.1909 ± .0144 •	.0916 ± .0022 •	.0892 ± .0050 •	.0919 ± .0022 •	**.0887 ± .0013**
Human Gene	.0535 ± .0007	.1826 ± .0198 •	.0540 ± .0040 •	.0602 ± .0009 •	.0647 ± .0007 •	**.0532 ± .0007**
LALOT W/ T/ L	10/ 5/ 0	14/ 1/ 0	13/ 2/ 0	13/ 0/ 2	12/ 0/ 3	rank first 12/ 15

(b) Performance Measure: `Cosine` ↑

Dataset	IIS-LLD	PT-Bayes	PT-SVM	AA-BP	AA-KNN	LALOT
Image	.5154 ± .0036 ○	.4882 ± .0064 •	.3618 ± .0684 •	**.6261 ± .0520** ○	.6220 ± .0100 ○	.4908 ± .0027
JAFFE-5	.9229 ± .0035 •	.6331 ± .0303 •	.9155 ± .0053 •	.9080 ± .0096 •	.9122 ± .0072 •	**.9304 ± .0034**
JAFFE-6	.9306 ± .0048	.6561 ± .0229 •	.9274 ± .0056 •	.9203 ± .0084 •	.9124 ± .0035 •	**.9307 ± .0042**
Emotions	.6253 ± .0121 ○	.4352 ± .0361 •	.3893 ± .1232 •	**.6979 ± .0184** ○	.6892 ± .0194 ○	.5513 ± .0065
Yeast_cdc	.9872 ± .0004 •	.8336 ± .0099 •	.9927 ± .0003	.9598 ± .0038 •	.9920 ± .0002 •	**.9928 ± .0002**
Yeast_cold	.9836 ± .0004 •	.8745 ± .0112 •	.9863 ± .0013 •	.9857 ± .0012 •	.9868 ± .0005 •	**.9873 ± .0005**
Yeast_diau	.9822 ± .0004 •	.8435 ± .0131 •	.9836 ± .0014 •	.9800 ± .0017 •	**.9860 ± .0003** ○	.9853 ± .0003
Yeast_dtt	.9892 ± .0004 •	.8785 ± .0149 •	.9935 ± .0004	.9910 ± .0010 •	.9931 ± .0002 •	**.9938 ± .0002**
Yeast_elu	.9877 ± .0003 •	.8359 ± .0089 •	.9932 ± .0003	.9643 ± .0029 •	.9929 ± .0002 •	**.9935 ± .0001**
Yeast_heat	.9813 ± .0003 •	.8468 ± .0065 •	.9868 ± .0007 •	.9810 ± .0010 •	.9860 ± .0002 •	**.9872 ± .0003**
Yeast_spo	.9715 ± .0007 •	.8503 ± .0094 •	.9728 ± .0020 •	.9698 ± .0035 •	.9720 ± .0007 •	**.9746 ± .0008**
Yeast_spo5	.9714 ± .0011 •	.8832 ± .0120 •	.9734 ± .0013 •	.9735 ± .0010 •	.9709 ± .0009 •	**.9741 ± .0007**
Yeast_spoem	.9762 ± .0008 •	.9132 ± .0099 •	.9750 ± .0024 •	- •	.9755 ± .0011 •	**.9771 ± .0007**
Human Gene	.8332 ± .0018	.4597 ± .0403 •	.8320 ± .0110	.7205 ± .0051 •	.7694 ± .0021 •	**.8333 ± .0018**
LALOT W/ T/ L	11/ 2/ 2	15/ 0/ 0	11/ 4/ 0	13/ 0/ 2	12/ 0/ 3	rank first 12/ 15

(ii) label correlations exploration: we will examine whether the obtained label correlations are reasonable.

Label Distribution Learning

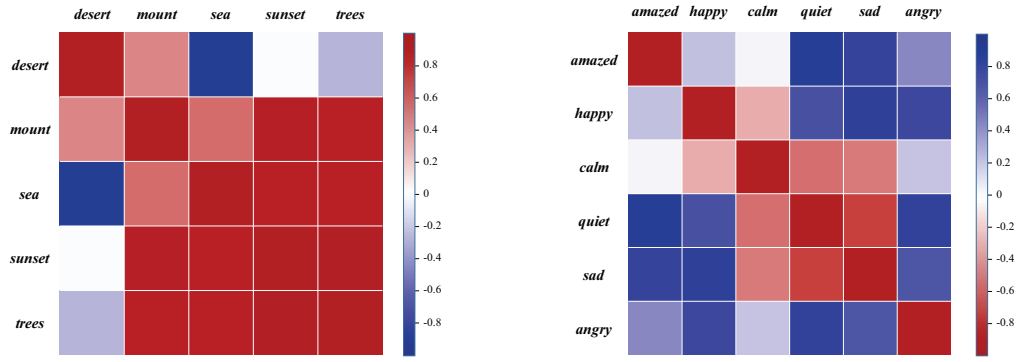
In this part, we will evaluate the proposed method on 15 real-world datasets with five state-of-the-art label distribution learning approaches over six different measurements.

Datasets The 15 datasets cover fields of biological information classification, natural scene recognition, emotional analysis and so on. Due to the page limitation, we only present the brief statistics of the datasets in Table 1.

Baselines As mentioned earlier, there are mainly three categories approaches. We compare the proposed LALOT to five state-of-the-art LDL algorithms, including two problem transformation methods PT-Bayes and PT-SVM (Geng 2016), two algothim adaptation methods AA-KNN and AA-BP (Geng 2016), and a specialized algorithms maximizing entropy IIS-LLD (Geng, Yin, and Zhou 2013).

Evaluation Six different measurements are used to evaluate the performance for LDL tasks. They can be divided into two groups, one type is to measure distance of two vectors including `Chebyshev`, `Clark`, `Canberra` and `KL divergence`. Obviously, these measurements are the lower the better. The other measurements measure similarity including `Cosine` and `Intersection`, which are the higher the better.

Parameter Settings For LALOT, there are two parameters. The first one is the trade-off parameter C, and the other is the entropic regularization coefficient λ. They are chosen by 10-fold cross-validation with random splitting 70% for



(a) *Image* Dataset (b) *Emotions* Dataset

Figure 1: Illustration of learned label correlations for different datasets, and the value has been scaled in [-1,1]. Red color indicates a positive correlation, and blue one indicates a negative correlation.

training and 30% for testing. One more attention is that λ should be chosen starting with a small value. Specifically, λ should small enough to make sure $\lambda\|M\|_\infty \leq 200$ because of the insufficient numerical precision, under which circumstance the Sinkhorns algorithm could blow up due to a too large λ value (Cuturi 2013).

Results There are actually six different measurements, we only present results w.r.t. `Chebyshev` (the lower the better) and `Cosine` (the higher the better) in Table 2. Performance results in terms of other measurements are similar and therefore omitted.

Form Table 2, we can see that in all the different measurements, LALOT outperforms the baselines, the results are expected since our approach could explicitly exploit label correlations to enhance learning performance. In a total of 15 datasets, the number of labels vary from 2 to at most 174. LALOT achieves the best among all approaches in 12 over 15 datasets. Also, in other 3 datasets, it ranks the second or the third. The reason LALOT behaves poorly on Image and Emotion might be that the distributions on these two datasets are discrete, which may cause a negative result since the transportation learned by OT is continuous in general. It is also noteworthy to mention that the behavior of LALOT is relatively consistent on different measurements. This validates the effectiveness of proposed LALOT.

Label Correlations Exploration

As mentioned before, one of our method's advantages is that it could learn label correlations explicitly. In this part, we examine the effectiveness of proposed algorithm in label correlations exploration. The explorations are conducted on two real-world datasets *Image* and *Emotions* (Zhou and Zhang 2007). Image dataset contains five labels: *desert*, *mountains*, *sea*, *sunset* and *trees*. Emotions dataset contains six different emotions as labels: *amazed*, *happy*, *calm*, *quiet* and *sad*.

The ground metric learned by LALOT are shown in Figure 1, and we scale the original value in cost matrix into [-1,1]. Red color indicates a positive correlation, and blue one indicates a negative correlation.

We can see that the learned pairwise cost accords with intuitions. Take a few examples, in Figure 1(a), the cost between (*desert*, *sea*) ranks the top indicating a very small correlation, and this is reasonable since when there is a *desert* occurring in an image, it is unlikely to find *sea* in the image generally. It is also noteworthy to mention that the label correlations discovered by LALOT are highly similar to the results reported in (Huang, Yu, and Zhou 2012). Specifically, four of top 5 most related label pairs discovered by LALOT also occur in theirs, and four of bottom 5 most related ones coincide with theirs. In Figure 1(b), the cost between (*amazed*, *calm*) and (*happy*, *sad*) rank the top and cost between (*quiet*, *sad*) and (*quiet*, *calm*) are very small. All above accord with our knowledge of emotional relationships, like that in Plutchiks theory (Plutchik 1980).

Conclusion

In this paper, a novel approach called LALOT is proposed to learn the label distribution based on optimal transport theory, and cast label correlations exploration as a ground metric learning problem. LALOT can avoid a costly projection to metric space by kernel biased regularization. During the optimization, an entropic regularization approach is adopted to approximate OT distance to speed up the computation.

Besides, we provide perhaps the first data-dependent risk bound analysis for label distribution learning, especially for the Sinkhorn distance, which is commonly used to approximately solve the optimal transport problem. Moreover, the effectiveness of LALOT is validated in experimental parts with state-of-the-art methods.

As for future work, one interesting issue is to parallelize the algorithm since the Sinkhorn distance can be easily vectorized and generalized to multiple label distributions. Another interesting one related to theoretical analysis is how to use novel mathematical tools to obtain a fast rate risk analysis for Sinkhorn distance based ERM problems.

Acknowledgement

Authors want to thank reviewers for helpful comments, and thank Han-Jia Ye for helpful discussions.

References

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.

Bertsimas, D., and Tsitsiklis, J. N. 1997. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA.

Bogachev, V. I., and Kolesnikov, A. V. 2012. The monge-kantorovich problem: achievements, connections, and perspectives. *Russian Mathematical Surveys* 67(5):785.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2016. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99):1–1.

Cuturi, M., and Avis, D. 2014. Ground metric learning. *Journal of Machine Learning Research* 15(1):533–564.

Cuturi, M., and Doucet, A. 2014. Fast computation of wasserstein barycenters. In *ICML*, 685–693.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2292–2300.

Frogner, C.; Zhang, C.; Mobahi, H.; Araya-Polo, M.; and Poggio, T. A. 2015. Learning with a wasserstein loss. In *NIPS*, 2053–2061.

Gao, B.-B.; Xing, C.; Xie, C.-W.; Wu, J.; and Geng, X. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26(6):2825–2838.

Geng, X.; Yin, C.; and Zhou, Z.-H. 2013. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(10):2401–2412.

Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28(7):1734–1748.

Huang, S.-J.; Yu, Y.; and Zhou, Z.-H. 2012. Multi-label hypothesis reuse. In *KDD*, 525–533.

Kedem, D.; Tyree, S.; Weinberger, K. Q.; Sha, F.; and Lanckriet, G. R. G. 2012. Non-linear metric learning. In *NIPS*, 2582–2590.

Koltchinskii, V. 2011. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Kong, S. G., and Mbouna, R. O. 2015. Head pose estimation from a 2d face image using 3d face morphing with depth parameters. *IEEE Transactions on Image Processing* 24(6):1801–1808.

Kuzborskij, I., and Orabona, F. 2017. Fast rates by transferring from auxiliary hypotheses. *Machine Learning* 106(2):171–195.

Maurer, A. 2016. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory - 27th International Conference, ALT 2016*, 3–17.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of machine learning*. MIT press.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. *Theories of Emotion* 1(3-31):4.

Qian, W.; Hong, B.; Cai, D.; He, X.; and Li, X. 2016. Non-negative matrix factorization with sinkhorn distance. In *IJCAI*, 1960–1966.

Rolet, A.; Cuturi, M.; and Peyré, G. 2016. Fast dictionary learning with a smoothed wasserstein loss. In *AISTATS*, 630–638.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* 40(2):99–121.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Xu, M., and Zhou, Z.-H. 2017. Incomplete label distribution learning. In *IJCAI*, 3175–3181.

Ye, J.; Wu, P.; Wang, J. Z.; and Li, J. 2017. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing* 65(9):2317–2332.

Zhang, M.-L., and Zhou, Z.-H. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7):2038–2048.

Zhang, M.-L., and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.

Zhou, Z.-H., and Zhang, M.-L. 2007. Multi-instance multi-label learning with application to scene classification. In *NIPS*, 1609–1616.

Zhou, Z.-H.; Zhang, M.-L.; Huang, S.-J.; and Li, Y.-F. 2012. Multi-instance multi-label learning. *Artificial Intelligence* 176(1):2291–2320.

Zhou, D.; Zhang, X.; Zhou, Y.; Zhao, Q.; and Geng, X. 2016. Emotion distribution learning from texts. In *EMNLP*, 638–647.

Zhou, Y.; Xue, H.; and Geng, X. 2015. Emotion distribution recognition from facial expressions. In *ACMMM*, 1247–1250.