Efficient robust optimal transport: formulations and algorithms

Pratik Jawanpuria* N T V Satya Dev† Bamdev Mishra*

Abstract

The problem of robust optimal transport (OT) aims at recovering the best transport plan with respect to the worst possible cost function. In this work, we study novel robust OT formulations where the cost function is parameterized by a symmetric positive semi-definite Mahalanobis metric. In particular, we study several different regularizations on the Mahalanobis metric – element-wise p-norm, KL-divergence, or doubly-stochastic constraint – and show that the resulting optimization formulations can be considerably simplified by exploiting the problem structure. For large-scale applications, we additionally propose a suitable low-dimensional decomposition of the Mahalanobis metric for the studied robust OT problems. Overall, we view the robust OT (min-max) optimization problems as non-linear OT (minimization) problems, which we solve using a Frank-Wolfe algorithm. We discuss the use of robust OT distance as a loss function in multi-class/multi-label classification problems. Empirical results on several real-world tag prediction and multi-class datasets show the benefit of our modeling approach.

1 Introduction

Optimal transport (Peyré & Cuturi, 2019) has become a popular tool in diverse machine learning applications such as domain adaptation (Courty et al., 2017), multi-task learning (Janati et al., 2017), natural language processing (Alvarez-Melis & Jaakkola, 2018), computer vision (Rubner et al., 2000), classification (Frogner et al., 2015), generative model training (Genevay et al., 2018; Arjovsky et al., 2017), to name a few. The optimal transport (OT) metric between two probability measures, also known as the Wasserstein distance or the earth mover's distance (EMD), defines a geometry over the space of probability measures (Villani, 2009) and evaluates the minimal amount of work required to transform one measure into another with respect to the ground cost function.

Given two probability measures μ_1 and μ_2 over metric spaces $\mathcal{S}$ and $\mathcal{T}$, respectively, the optimal transport problem due to Kantorovich (1942) aims at finding a transport plan γ as a solution to the following problem:

$$\mathrm{W}_c(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{S}\times\mathcal{T}} c(\mathbf{s}, \mathbf{t}) d\gamma(\mathbf{s}, \mathbf{t}), \tag{1}$$

where $\Pi(\mu_1, \mu_2)$ is the set of joint distributions with marginals μ_1 and μ_2 and $c : \mathcal{S} \times \mathcal{T} \to \mathbb{R}_+ : (\mathbf{s}, \mathbf{t}) \to c(\mathbf{s}, \mathbf{t})$ represents the transportation cost function. We obtain the popular 2-Wasserstein

*Microsoft India. Email: {pratik.jawanpuria, bamdevm}@microsoft.com.
†Vayve Technologies. Email: tvsatyadev@gmail.com.

distance by setting the cost function as the squared Euclidean function, i.e., $c(\mathbf{s}, \mathbf{t}) = \|\mathbf{s} - \mathbf{t}\|^2$ when $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$. The 2-Wasserstein distance can be reformulated as follows (Paty & Cuturi, 2019):

$$\mathrm{W}_2^2(\mu_1, \mu_2) = \min_{\gamma \in \Pi(\mu_1, \mu_2)} \langle \mathbf{V}_\gamma, \mathbf{I} \rangle \,, \tag{2}$$

where $\mathbf{V}_\gamma \coloneqq \int_{\mathcal{S}\times\mathcal{T}} (\mathbf{s} - \mathbf{t})(\mathbf{s} - \mathbf{t})^\top d\gamma(\mathbf{s}, \mathbf{t})$ and $\mathbf{I}$ is the identity matrix.

Robust optimal transport

Recent works (Paty & Cuturi, 2019; Kolouri et al., 2019; Deshpande et al., 2019) have proposed variants of the Wasserstein distance that aim at maximizing the OT distance between two probability measures in a projected low-dimensional space, which may be viewed as instances of robust OT distance. In general, the robust OT distance aims at maximizing the minimal transport cost over a set of ground cost functions.

Paty & Cuturi (2019) propose a robust variant of the W_2^2 distance, termed as the Subspace Robust Wasserstein (SRW) distance, as follows:

$$\mathrm{SRW}_k^2(\mu_1, \mu_2) = \max_{\mathbf{M} \in \mathcal{M}} \min_{\gamma \in \Pi(\mu_1, \mu_2)} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle \,, \tag{3}$$

where the domain $\mathcal{M}$ is defined as $\mathcal{M} = \{\mathbf{M} : \mathbf{0} \preceq \mathbf{M} \preceq \mathbf{I} \text{ and } \mathrm{trace}(\mathbf{M}) = k\}$. It should be noted that $\langle \mathbf{V}_\gamma, \mathbf{M} \rangle = \int_{\mathcal{S}\times\mathcal{T}} c_{\mathbf{M}}(\mathbf{s}, \mathbf{t}) d\gamma(\mathbf{s}, \mathbf{t})$, where $c_{\mathbf{M}}(\mathbf{s}, \mathbf{t}) = (\mathbf{s} - \mathbf{t})^\top \mathbf{M} (\mathbf{s} - \mathbf{t})$ is a cost function parameterized by a symmetric positive semi-definite matrix or a Mahalanobis metric $\mathbf{M}$ of size $d \times d$.

Dhouib et al. (2020) also study the form (3) of Mahalanobis metric parameterized cost functions in the robust OT setting, but with the domain $\mathcal{M}$ defined as $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0} \text{ and } \|\mathbf{M}\|_{*p} = 1\}$, where $\|\cdot\|_{*p}$ denotes the Schatten p-norm regularizer, i.e., $\|\mathbf{M}\|_{*p} \coloneqq (\sum_i \sigma_i(\mathbf{M})^p)^{\frac{1}{p}}$. Here, $\sigma_i(\mathbf{M})$ denotes the i-th largest eigenvalue of $\mathbf{M}$.

Both (Paty & Cuturi, 2019; Dhouib et al., 2020) pose their Mahalanobis metric parameterized robust optimal transport problems as optimization problems over the metric $\mathbf{M}$. This involves satisfying the positive semi-definite constraint at every iteration, which typically requires costly eigenvalue decomposition operations costing $O(d^3)$.

Contributions

In this paper, we focus on the robust OT distance formulations arising from Mahalanobis metric parameterized cost functions and discuss several novel regularizations and algorithmic approaches to use them in learning problems. Our main contributions are the following.

- We discuss three families of regularizers on $\mathbf{M}$: entry-wise p-norm for $p \in (1, 2]$, KL-divergence based regularization, and the doubly-stochastic regularization. We show that for those regularizers, enforcement of the symmetric positive semi-definite constraint is not needed as the problem structure *implicitly* learns a symmetric positive semi-definite matrix at optimality. The implications are two fold. First, the computational cost of computing the robust OT distance is $O(d^2)$ as against $O(d^3)$ (when using the existing regularizers). Second, our optimization methodology for computing the robust OT distance simplifies.

- To further reduce the computational burden of the robust OT distance computation, we propose a novel r-dimensional decomposition of the Mahalanobis metric for the studied robust OT problems, resulting in the per-iteration computational cost of $O(r^2)$, where $r \ll d$. The parameter r provides an effective trade-off between computational efficiency and accuracy.
- We discuss how to use the robust distance as a loss in multi-class/multi-label classification problems.

The outline of the paper is as follows. Section 2 discusses the proposed regularizations. The novel decomposition of the metric is discussed in Section 3. We formulate the robust OT distance problem as a non-linear OT (minimization) problem in γ. In Section 4, this interpretation allows to develop a Frank-Wolfe algorithm. The setup of using the robust OT distance in learning problems is discussed in Section 5. In Section 6, we show the good performance of our modeling approach in several real-world prediction datasets: Animals with attributes, MNIST, and Flickr.

2 Novel formulations for robust optimal transport

In this section, we propose novel formulations of the Mahalanobis metric parameterized robust optimal transport problem, which we rewrite as

$$\mathrm{W}_{\mathrm{ROT}}(\mu_1, \mu_2) \coloneqq \min_{\gamma \in \Pi(\mu_1, \mu_2)} f(\gamma), \tag{4}$$

where the function $f : \mathbb{R}^{m \times n} \to \mathbb{R} : \gamma \mapsto f(\gamma)$ is defined as

$$f(\gamma) \coloneqq \max_{\mathbf{M} \in \mathcal{M}} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle . \tag{5}$$

Here, $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0} \text{ and } \Omega(\mathbf{M}) \leq 1\}$ and $\Omega(\cdot)$ is a convex regularizer on the set of positive semi-definite matrices. It should be noted that (4) is a convex optimization problem. Moreover, by the application of Sion-Kakutani min-max theorem (Sion, 1958), Problem (4) can be shown to be equivalent to its dual max-min problem: $\max_{\mathbf{M} \in \mathcal{M}} \min_{\gamma \in \Pi(\mu_1, \mu_2)} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle$.

First-order methods for solving (4) requires computing the (sub-)gradient $\nabla f(\gamma)$, which can be obtained in terms of an optimal solution $\mathbf{M}^*(\gamma)$ of (5) by using the Danskin's theorem (Bertsekas, 1995). Since (5) involves the positive semi-definite constraint, computing $\mathbf{M}^*(\gamma)$ usually involves costly eigenvalue decomposition (or equivalent operations) of $d \times d$ matrices. However, for the proposed family of regularizers $\Omega(\cdot)$, we show that one can drop the symmetric positive definite constraint in (5) as the optimal solution $\mathbf{M}^*(\gamma)$ of the resulting problem automatically satisfies $\mathbf{M}^*(\gamma) \succeq \mathbf{0}$. This considerably simplifies our optimization methodology for the proposed formulations. We discuss the proposed regularizers and the corresponding $\mathbf{M}^*(\gamma)$ in the following sections.

2.1 Element-wise p-norm regularization on M

We consider $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0} \text{ and } \|\mathbf{M}\|_p \leq 1\}$ in (5), where $\|\mathbf{M}\|_p = (\sum_{ij} |\mathbf{M}_{ij}|^p)^{\frac{1}{p}}$ denotes the element-wise p-norm on the matrix $\mathbf{M}$.

It should be noted that $\mathbf{V}_\gamma$ is the second-order moment matrix of the displacements associated with a transport plan, i.e., for source and target distribution's samples $\{\mathbf{s}_i\}_{i=1}^m \in \mathbb{R}^d$ and $\{\mathbf{t}_j\}_{j=1}^n \in \mathbb{R}^d$, respectively, $\mathbf{V}_\gamma = \sum_{ij} (\mathbf{s}_i - \mathbf{t}_j)(\mathbf{s}_i - \mathbf{t}_j)^\top \gamma_{ij}$. Hence, the proposed regularization on the metric

$\mathbf{M}$ learns appropriate weights to the individual components of $\mathbf{V}_\gamma$. In contrast, the W_2^2 distance, defined in (2), has $\mathbf{M} = \mathbf{I}$, which enforces the first-order displacements to be uncorrelated and have unit variance. This may be a strong assumption in real-word applications.

The family of element-wise p-norm regularizers includes the popular Frobenius norm at $p = 2$. For p in between 1 and 2, the entry-wise p-norm regularization induces a sparse structure on the metric $\mathbf{M}$. A sparse Mahalanobis metric is useful for working with high dimensional features as it helps to avoid spurious correlations (Rosales & Fung, 2006; Qi et al., 2009). The following result provides an efficient reformulation of the robust OT problem (4) with the above defined $\mathcal{M}$ for a subset of the element-wise p-norm regularizers on $\mathbf{M}$.

Theorem 2.1. *Let $k \in \mathbb{N}$, $p = \frac{2k}{2k-1}$, and $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0}$ and $\|\mathbf{M}\|_p \leq 1\}$. Consider the optimization problem*

$$\mathrm{W_P}(\mu_1, \mu_2) \coloneqq \min_{\gamma \in \Pi(\mu_1, \mu_2)} \|\mathbf{V}_\gamma\|_{2k}, \tag{6}$$

where $\mathbf{V}_\gamma = \sum_{ij} (\mathbf{s}_i - \mathbf{t}_j)(\mathbf{s}_i - \mathbf{t}_j)^\top \gamma_{ij}$. Then, the following statements hold.

1. *Problem (6) is a dual of (4) and the objectives of (4) and (6) are equal at optimality.*
2. *For a given $\gamma \in \Pi(\mu_1, \mu_2)$, the optimal solution of (5) is $\mathbf{M}^*(\gamma) = \|\mathbf{V}_\gamma\|_{2k}^{1-2k} (\mathbf{V}_\gamma)^{\circ(2k-1)}$, where $\mathbf{A}^{\circ(k)}$ denotes the k-th Hadamard power of a matrix $\mathbf{A}$, i.e., $\mathbf{A}^{\circ(k)}(s,t) = \mathbf{A}(s,t)^k$.*

The proof of the above theorem involves a result derived in the context of multi-task learning in (Jawanpuria et al., 2015). The proof details are in Section A.1. It should be observed that for a given γ, the optimal $\mathbf{M}^*(\gamma)$ is an element-wise function of the matrix $\mathbf{V}_\gamma$. An implication is that the computation of $\mathbf{M}^*(\gamma)$ costs $O(d^2)$.

An optimal solution γ^* of (6) is also an optimal solution the problem:

$$\mathrm{W_P^{2k}} = \min_{\gamma \in \Pi(\mu_1, \mu_2)} \|\mathbf{V}_\gamma\|_{2k}^{2k}. \tag{7}$$

From an optimization perspective the gradient expression for (7) is simpler than that of (6).

The following result establishes a general bound on the $\mathrm{W_P}$ distance in terms of the W_2^2 (2-Wasserstein) distance as defined in (2).

Corollary 2.2. *Let $k \in \mathbb{N}$, $p = \frac{2k}{2k-1}$, $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0}; \|\mathbf{M}\|_p \leq 1\}$, and the $\mathrm{W_P}(\mu_1, \mu_2)$ distance as defined in (6). Then,*

$$\frac{1}{d^{\frac{1}{p}}} \mathrm{W}_2^2(\mu_1, \mu_2) \leq \mathrm{W_P}(\mu_1, \mu_2) \leq \mathrm{W}_2^2(\mu_1, \mu_2).$$

The proof of Corollary 2.2 is in Section A.2.

2.2 KL-divergence regularization on M

We next consider the generalized KL-divergence regularization on the metric $\mathbf{M}$, i.e., $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0}; \mathrm{D_{KL}}(\mathbf{M}, \mathbf{M}_0) \leq 1\}$, where $\mathrm{D_{KL}}(\mathbf{M}, \mathbf{M}_0) = \sum_{s,t} \mathbf{M}_{st}(\ln(\mathbf{M}_{st}/\mathbf{M}_{0st})) - (\mathbf{M}_{st} - \mathbf{M}_{0st})$ denotes the Bregman distance, with negative entropy as the distance-generating function, between the matrices $\mathbf{M}$ and $\mathbf{M}_0 \succeq \mathbf{0}$. Here, $\mathbf{M}_0$ is a given symmetric positive semi-definite matrix, which may be useful in introducing a prior domain knowledge (e.g., a block diagonal matrix $\mathbf{M}_0$ ensures

grouping of features). The function $f(\gamma)$ in (5) with the above defined $\mathcal{M}$ may be expressed in the Tikhonov form as

$$f(\gamma) = \max_{\mathbf{M} \succeq \mathbf{0}} \ \langle \mathbf{V}_\gamma, \mathbf{M} \rangle - \lambda_{\mathbf{M}} \mathrm{D}_{\mathrm{KL}}(\mathbf{M}, \mathbf{M}_0), \tag{8}$$

where $\lambda_{\mathbf{M}} > 0$ is a regularization parameter. The following result provides an efficient reformulation of the KL-divergence regularized robust OT problem.

Theorem 2.3. *The following problem is an equivalent dual of the convex problem (4) where $f(\gamma)$ is as defined in (8):*

$$\mathrm{W}_{\mathrm{KL}}(\mu_1, \mu_2) = \min_{\gamma \in \Pi(\mu_1, \mu_2)} \lambda_{\mathbf{M}} \mathbf{1}^\top (\mathbf{M}_0 \odot e^{\circ(\mathbf{V}_\gamma / \lambda_{\mathbf{M}})} - \mathbf{M}_0) \mathbf{1}, \tag{9}$$

where $e^{\circ \mathbf{A}}$ denotes the Hadamard exponential (element-wise exponential) of a matrix $\mathbf{A}$. For a given $\gamma \in \Pi(\mu_1, \mu_2)$, the optimal solution of (8) is $\mathbf{M}^*(\gamma) = \mathbf{M}_0 \odot e^{\circ(\mathbf{V}_\gamma / \lambda_{\mathbf{M}})}$.

We next analyze the doubly-stochastic structure on $\mathbf{M}$.

2.3 Doubly-stochastic regularization on M

We consider the robust OT problem (4) in which the metric $\mathbf{M}$ is also doubly-stochastic, i.e.,

$$f(\gamma) = \max_{\mathbf{M} \in \mathcal{M}} \ \langle \mathbf{V}_\gamma, \mathbf{M} \rangle \,, \tag{10}$$

where $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0}, \mathbf{M} \geq \mathbf{0}, \text{ and } \mathbf{M}\mathbf{1} = \mathbf{1}\}$. Learning of doubly-stochastic Mahalanobis metric is of independent interest in applications such as graph clustering and community detection (Zass & Shashua, 2006; Arora et al., 2011; Wang et al., 2016; Douik & Hassibi, 2018, 2019).

In general, optimization over the set of positive semi-definite stochastic matrices is non trivial and computationally challenging. By exploiting the problem structure, however, we show that (10) can be solved efficiently by adding a small KL-divergence regularization. To this end, we propose to solve for

$$\tilde{f}(\gamma) = \max_{\mathbf{M} \in \mathcal{M}} \ \langle \mathbf{V}_\gamma, \mathbf{M} \rangle - \lambda_{\mathbf{M}} \mathrm{D}_{\mathrm{KL}}(\mathbf{M}, \mathbf{M}_0), \tag{11}$$

where $\lambda_{\mathbf{M}} > 0$ is a small regularization parameter and $\mathbf{M}_0$ is a symmetric positive semi-definite prior which is also element-wise positive. Our next result discusses the solution of (11).

Theorem 2.4. *For a given $\gamma \in \Pi(\mu_1, \mu_2)$, the optimal solution of (11) has the form $\mathbf{M}^*(\gamma) = \mathbf{D}\left(\mathbf{M}_0 \odot e^{\circ(\mathbf{V}_\gamma / \lambda_{\mathbf{M}})}\right)\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with positive entries.*

The matrix $\mathbf{D}$ in Theorem 2.4 is efficiently computed using the Sinkhorn algorithm (Cuturi, 2013).

3 Robust optimal transport with low-dimensional metric M

Section 2 discusses several regularizations on the Mahalanobis metric $\mathbf{M}$ that lead to efficient computation of $\mathbf{M}^*(\gamma)$, i.e., the solution to (5). For a given $\mathbf{V}_\gamma$, computation of $\mathbf{M}^*(\gamma)$, for regularizations considered in Section 2, requires $O(d^2)$, which though linear in the size of $\mathbf{M}$, may be prohibitive for high-dimensional data.

In this section, we discuss a particular dimensionality reduction technique that addresses the computational issue. To this end, we propose a novel decomposition of the Mahalanobis metric $\mathbf{M}$ as

$$\mathbf{M} = \mathbf{B} \otimes \mathbf{I}_{d_1}, \tag{12}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{I}_{d_1}$ denotes the identity matrix of size $d_1 \times d_1$, and $\mathbf{B} \succeq \mathbf{0}$ is a $r \times r$ symmetric positive semi-definite matrix such that $d = d_1 r$, where $r \ll d$. The decomposition (12) induces the following reformulation of the objective in (5) as

$$\begin{aligned} \langle \mathbf{V}_\gamma, \mathbf{B} \otimes \mathbf{I}_{d_1} \rangle &= \left\langle \textstyle\sum_{ij} \gamma_{ij} (\mathbf{s}_i - \mathbf{t}_j)(\mathbf{s}_i - \mathbf{t}_j)^\top, \mathbf{B} \otimes \mathbf{I}_{d_1} \right\rangle \\ &= \left\langle \textstyle\sum_{ij} \gamma_{ij} (\mathbf{S}_i - \mathbf{T}_j)^\top (\mathbf{S}_i - \mathbf{T}_j), \mathbf{B} \right\rangle \\ &= \langle \mathbf{U}_\gamma, \mathbf{B} \rangle , \end{aligned} \tag{13}$$

where $\mathbf{S}_i$ and $\mathbf{T}_j$ are $d_1 \times r$ matrices obtained by reshaping the vectors $\mathbf{s}_i$ and $\mathbf{t}_j$, respectively.

We observe that the proposed decomposition of the Mahalanobis metric (12) divides the d features into r groups, each with d_1 input features. Based on (13), the symmetric positive semi-definite matrix $\mathbf{B}$ may be viewed as a Mahalanobis metric over the feature groups. In addition, it can be shown that any proposed regularization on the metric $\mathbf{M}$ (in Section 2) transforms into an equivalent regularization on the "group" metric $\mathbf{B}$.

The function $f(\gamma)$, in the robust optimal transport problem (4), is equivalently re-written as

$$f(\gamma) = \max_{\mathbf{B} \in \mathcal{B}} \ \langle \mathbf{U}_\gamma, \mathbf{B} \rangle , \tag{14}$$

where $\mathcal{B} = \{\mathbf{B} : \mathbf{B} \succeq \mathbf{0} \text{ and } \Omega(\mathbf{B}) \leq 1\}$. For the regularizers discussed in Section 2, the computation of the solution $\mathbf{B}^*(\gamma)$ of (14) costs $O(r^2)$.

4 The Frank-Wolfe algorithm for (4)

The formulations proposed in Section 2 are expressed as minimization of a non-linear convex function $f : \Pi \to \mathbb{R} : \gamma \mapsto f(\gamma)$ over $\Pi(\mu_1, \mu_2)$. Here, the objective function f encapsulates the Mahalanobis metric $\mathbf{M}$ and is more generically written as $f(\gamma) \coloneqq \max_{\mathbf{M} \in \mathcal{M}} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle$.

A popular way to solve a convex constrained optimization problem (4) is with the Frank-Wolfe (FW) algorithm, which is also known as the conditional gradient algorithm. It requires solving a constrained linear minimization sub-problem (LMO) at every iteration. For many convex constraints, the LMOs are often easy to solve, thereby making the FW algorithm an appealing choice in practice (Jaggi, 2013).

The proposed algorithm for (4) is shown in Algorithm 1. The LMO step boils down to solving the optimal transport problem (1), where the cost matrix is replaced by $\nabla_\gamma f$. When regularized with an entropy regularization term, the LMO step admits a computationally efficient solution using the Sinkhorn algorithm (Cuturi, 2013).

We now develop the expression of $\nabla_\gamma f$ and discuss its computational cost. We begin by noting that $\mathbf{V}_\gamma = \mathbf{Z} \text{Diag}(\text{vec}(\gamma)) \mathbf{Z}^\top$, where $\mathbf{Z}$ is a $d \times mn$ matrix with (i,j)-th column as $(\mathbf{s}_i - \mathbf{t}_j)$, $\text{Diag}(\cdot)$ acts on a vector and outputs the corresponding diagonal matrix, and $\text{vec}(\cdot)$ vectorizes a matrix in the column-major order.

Algorithm 1 Proposed FW algorithm for (4)

Input: Source distribution's samples $\{\mathbf{s}_i\}_{i=1}^m \in \mathbb{R}^d$ and target distribution's samples $\{\mathbf{t}_j\}_{j=1}^n \in \mathbb{R}^d$. Initialize $\gamma_0 \in \Pi(\mu_1, \mu_2)$.
for $t = 0 \dots T$ **do**
  Compute $\nabla f_\gamma(\gamma_t)$ using Lemma 4.1.
  LMO step: Compute $\hat{\gamma}_t \coloneqq \underset{\beta \in \Pi(\mu_1,\mu_2)}{\arg\min} \langle \beta, \nabla f_\gamma(\gamma_t) \rangle$.
  Update $\gamma_{t+1} = (1-\theta)\gamma_t + \theta\hat{\gamma}_t$ for $\theta = \frac{2}{t+2}$.
end for
Output: γ^* and $\mathbf{M}^* = \mathbf{M}^*(\gamma^*)$.

Lemma 4.1. *Let* $\mathbf{M}^*(\gamma_t)$ *be the solution to the problem* $\max_{\mathbf{M} \in \mathcal{M}} \langle \mathbf{V}_{\gamma_t}, \mathbf{M} \rangle$ *for a given* $\gamma_t \in \Pi(\mu_1, \mu_2)$. *Then, the gradient* $\nabla_\gamma f(\gamma_t)$ *of* f *in (4) with respect to* γ *has the expression*

$$\nabla_\gamma f(\gamma_t) = \text{vec}^{-1}(\text{diag}(\mathbf{Z}^\top \mathbf{M}^*(\gamma_t)\mathbf{Z})),$$

where $\text{diag}(\cdot)$ *extracts the diagonal (vector) of a square matrix and* vec^{-1} *reshapes a vector into a matrix.*

The proof of Lemma 4.1 is in Section A.5. Lemma 4.1 shows that the gradient $\nabla_\gamma f$ can be written by using $\mathbf{M}^*(\gamma)$. For the formulations presented in Section 2, the expressions for $\mathbf{M}^*(\gamma)$ can be computed efficiently. Overall, the computation of $\nabla_\gamma f$ costs $O(mnd^2)$.

For the cost function (14) obtained with the r-dimensional decomposition of $\mathbf{M}$, Lemma 4.1 can be appropriately modified. In this case, the cost of computation of $\nabla_\gamma f$ is $O(mnr^2)$.

5 Learning with robust optimal transport loss

Optimal transport based distances provide a natural metric on the output space, and hence, can be used to in learning problems. Specifically, Frogner et al. (2015) discuss the multi-label classification problem set up where they employ the Wasserstein distance as the loss function.

In this section, we discuss the use of the robust OT distance as a loss function in learning problems. We begin by explaining the learning problem setup with W_{ROT} as a loss function. We then discuss the qualitative nature of the proposed robust OT distance-based loss functions and present an efficient way to compute the gradient of the robust OT distance loss. Later in Section 6, we empirically demonstrate the effectiveness of using W_{ROT} as loss function in supervised multi-label/multi-class classification problems.

5.1 Problem setup

Consider the standard multi-label (or multi-class) problem over L labels (classes) and given supervised training instances $\{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^N$. Here, $\mathbf{x}_j \in \mathbb{R}^M$ and $\mathbf{y}_j \in \{0,1\}^L$. The prediction function of p-th label is given by the *softmax* function

$$h_p(\mathbf{x}; \mathbf{W}) = \frac{e^{\langle \mathbf{w}_p, \mathbf{x} \rangle}}{\sum_{l=1}^L e^{\langle \mathbf{w}_l, \mathbf{x} \rangle}}, \tag{15}$$

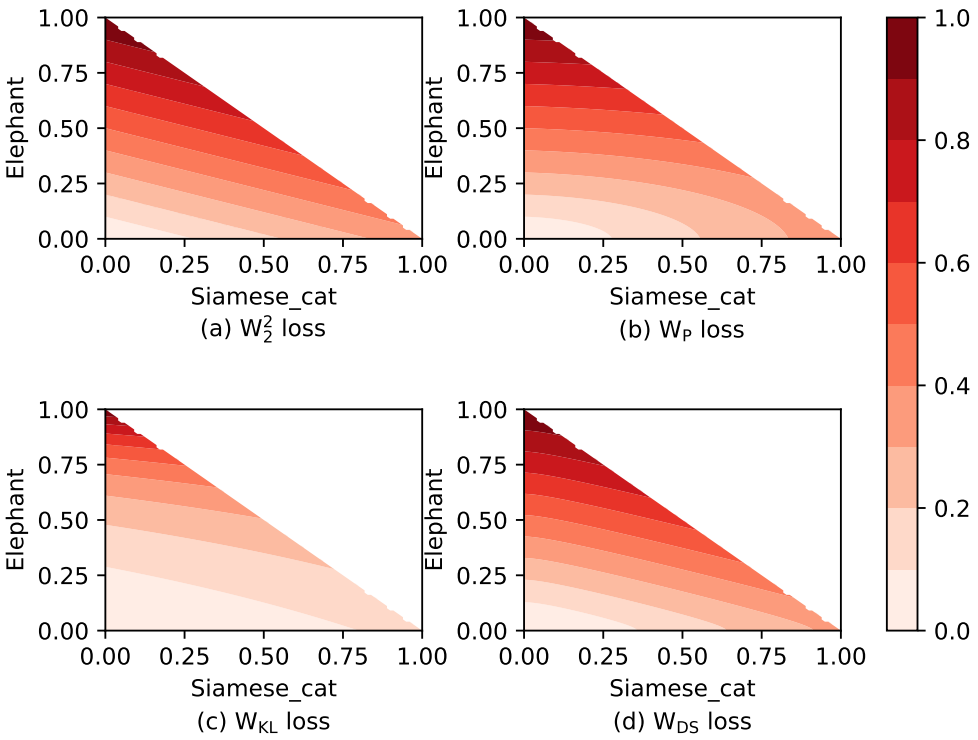

Figure 1: Contour plots of various OT distance based loss functions. The labels are {'Siamese_cat', 'Elephant', 'Persian_cat'}, with 'Persian_cat' as the true label. The plots show the relative loss incurred by various possible predictions of the form $\mathbf{h} = [x, y, 1-x-y]^\top$, where $0 \leq x, y \leq 1$ and $x + y \leq 1$. We observe linear contours for the W_2^2 loss function while non-linear contours for the proposed robust OT loss functions: $\mathrm{W_P}$ with $k = 1$, $\mathrm{W_{KL}}$, and $\mathrm{W_{DS}}$. The robust OT loss functions relatively penalize 'Siamese_cat' lesser than 'Elephant', i.e., more lighter color in the feasible $x > y$ regions in plots (b), (c), and (d) than in plot (a).

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]$ is the model parameter of the multi-label problem. The prediction function $h(\mathbf{x}; \mathbf{W})$ can be learned via the empirical risk minimization framework.

Frogner et al. (2015) use the Wasserstein distance as the loss function for multi-label classification problem as follows. For an input $\mathbf{x}_j$, the prediction function $h(\mathbf{x}_j; \mathbf{W})$ in Equation (15) may be viewed as a discrete probability distribution. Similarly, a $0-1$ binary ground-truth label vector $\mathbf{y}_j$ may be transformed into a discrete probability distribution by appropriate normalization: $\hat{\mathbf{y}}_j = \mathbf{y}_j / \mathbf{y}_j^\top \mathbf{1}$ (Frogner et al., 2015). Given a suitable ground cost metric between the labels, the Wasserstein distance is employed to measure the distance between the prediction $h(\mathbf{x}_j; \mathbf{W})$ and the ground-truth $\hat{\mathbf{y}}_j$. If the labels correspond to real-word entities, then a possible ground cost metric may be obtained from the word embedding vectors corresponding to the labels (Mikolov et al., 2013; Bojanowski et al., 2016).

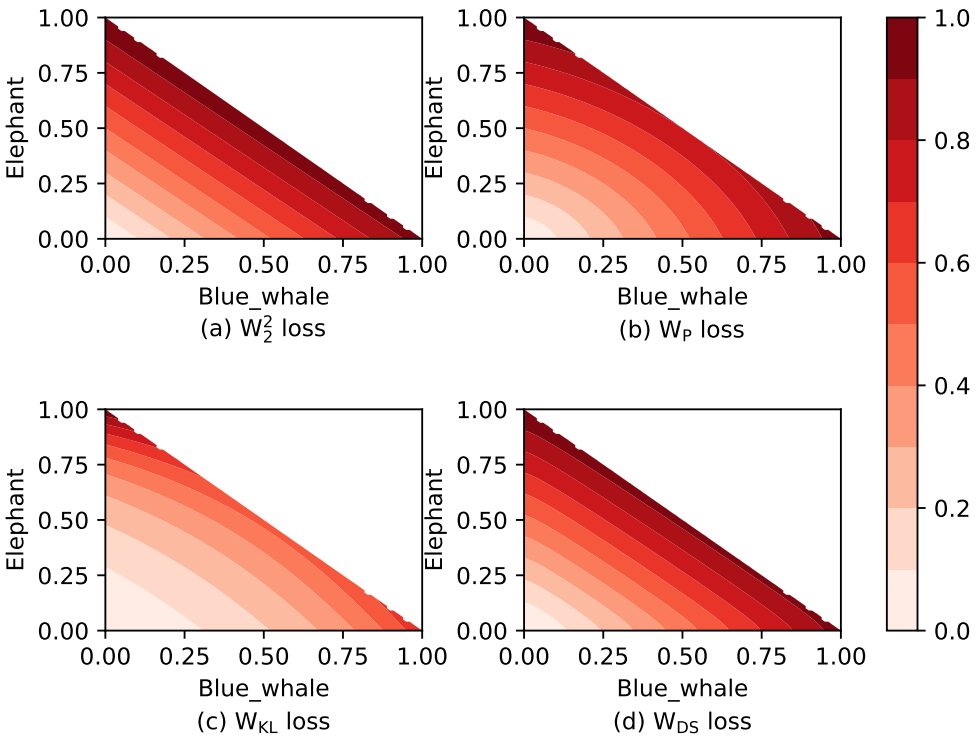

Figure 2: Contour plots of various OT distance based loss functions. The labels are {'Blue_whale', 'Elephant', 'Persian_cat'}, with 'Persian_cat' being the true label. All three classes are quite dissimilar from each other. Hence, all OT based loss functions relatively penalize both the incorrect classes similarly.

5.2 Multi-label learning with the $\mathrm{W}_{\mathrm{ROT}}$ loss

We propose to employ the robust OT distance-based loss in multi-label/multi-class problems. To this end, we solve the empirical risk minimization problem, i.e.,

$$\min_{\mathbf{W}\in\mathbb{R}^{M\times L}} \frac{1}{N}\sum_{j=1}^{N} \mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j;\mathbf{W}),\hat{\mathbf{y}}_j), \tag{16}$$

where $\mathrm{W}_{\mathrm{ROT}}$ is the robust OT distance-based function (4). Here, $\mathrm{W}_{\mathrm{ROT}}$ may be set to any of the discussed robust OT distance functions such as W_{P} (6), W_{KL} (9), and W_{DS} (11). As discussed, Frogner et al. (2015) employ $\mathrm{W}_p^p(h(\mathbf{x};\mathbf{W}),\hat{\mathbf{y}})$ as the loss function in (16).

We analyze the nature of the W_2^2-based loss function and the proposed $\mathrm{W}_{\mathrm{ROT}}$-based loss functions by viewing their contours plots for the three-class setting. We consider the labels as $\{A,B,C\}$. We consider label C as true label, i.e., $\hat{\mathbf{y}} = [0,0,1]^\top$, where the first dimension corresponds to label A and consider predictions of the form $\mathbf{h} = [a,b,1-a-b]$. Since h is obtained from the softmax function (15), we have $(a,b) \in \{(x,y) : 0 \le x,y \le 1 \text{ and } x+y \le 1\}$. The ground cost function is computed using the fastText word embeddings corresponding to the labels (Bojanowski et al., 2016).

We plot the contour maps of the $\mathrm{W}_2^2(\mathbf{h},\hat{\mathbf{y}})$ loss function and the proposed $\mathrm{W}_{\mathrm{P}}(\mathbf{h},\hat{\mathbf{y}})$, $\mathrm{W}_{\mathrm{KL}}(\mathbf{h},\hat{\mathbf{y}})$, and $\mathrm{W}_{\mathrm{DS}}(\mathbf{h},\hat{\mathbf{y}})$ loss functions, as (a,b) varies along the two-dimensional $X-Y$ plane. All the plots

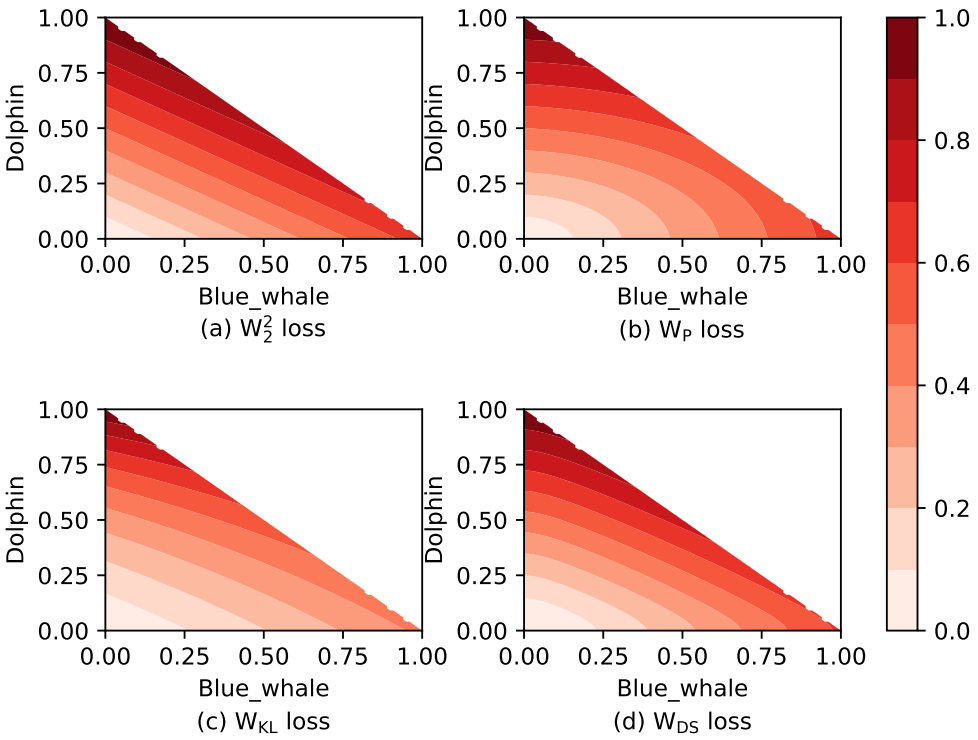

Figure 3: Contour plots of various OT distance based loss functions. The labels are {'Blue_whale', 'Dolphin', 'Humpback_whale'}, with 'Humpback_whale' as the true label. All three classes share similar characteristics but 'Blue_whale is more similar to 'Humpback_whale'.

are made to same scale by normalizing the highest value of the loss to 1.

In Figure 1, we consider the labels as $\{A,B,C\}$={'Siamese_cat', 'Elephant', 'Persian_cat'}. The first and the third classes in this setting are similar, and the OT distance based loss functions should exploit this relationship via the ground cost function. With 'Persian_cat' as the true class, we observe that the all four losses in Figure 1 penalize 'Elephant' more than 'Siamese_cat'. However, the contours of W_2^2 loss in Figure 1(a) are linear, while those of the proposed W_P with $k = 1$ Figure 1(b) are elliptical. On the other hand, the proposed W_KL and W_DS exhibit varying degree of non-linear contours (modeled by $\lambda_\mathbf{M}$ parameter). Overall, the proposed robust OT distance based loss functions penalize 'Siamese_cat' relatively lesser than the W_2^2 loss.

In Figure 2, we consider the labels as $\{A,B,C\}$={'Blue_cat', 'Elephant', 'Persian_cat'}. The two classes 'Blue_cat' and 'Elephant' are quite dissimilar to the true class 'Persian_cat'. We observe that while the W_2^2 loss exhibits similar linear contours as previously (but with a different slope), non-linearity in the proposed robust OT based losses allow them more freedom to adapt. We provide contour maps for another interesting setting, where all the three classes are similar, but one class is more closer to the true class. In Figure 3, we consider the labels as $\{A,B,C\}$={'Blue_whale', 'Dolphin', 'Humpback_whale'}.

Table 1: Generalization performance of various optimal transport based loss functions on the supervised multi-class/multi-label problems.

Method	r	Animals (AUC)	MNIST (AUC)	Flickr (AUC)	Flickr (mAP)
W_2^2 loss (Frogner et al., 2015)	–	0.794	0.848	0.649	0.0209
W_{P} loss with $k = 1$ (Eq. 6)	5	0.801	**0.959**	0.706	0.0481
	10	0.907	0.940	0.745	0.0625
	20	**0.908**	0.922	**0.768**	**0.0717**
W_{P} loss with $k = 2$ (Eq. 6)	5	0.890	0.804	0.760	0.0625
	10	0.878	0.702	0.742	0.0468
	20	0.874	0.603	0.737	0.0441
W_{KL} loss (Eq. 9)	5	0.794	0.867	0.648	0.0207
	10	0.749	0.870	0.649	0.0205
	20	0.794	0.779	0.649	0.0205
W_{DS} loss (Eq. 11)	5	**0.908**	0.927	0.724	0.0530
	10	0.878	0.873	0.668	0.0325
	20	0.845	0.813	0.628	0.0219

5.3 Optimizing the $\mathrm{W}_{\mathrm{ROT}}$ loss

We solve Problem (16) using the standard stochastic gradient descent (SGD) algorithm. In each iteration of the SGD algorithm, we pick a training instance $j = \{1, \ldots, N\}$ and update the parameter $\mathbf{W}$ along the negative of the gradient of the loss term $\mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j; \mathbf{W}), \hat{\mathbf{y}}_j))$ with respect to the model parameter $\mathbf{W}$.

We obtain it using the chain rule by computing $\nabla_{\mathbf{W}} h(\mathbf{x}_j; \mathbf{W})$ and $\nabla_{h(\mathbf{x};\mathbf{W})} \mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j; \mathbf{W}), \hat{\mathbf{y}}_j)$. While the expression for $\nabla_{\mathbf{W}} h(\mathbf{x}_j; \mathbf{W})$ is well studied, computing $\nabla_{h(\mathbf{x};\mathbf{W})} \mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j; \mathbf{W}), \hat{\mathbf{y}}_j)$ is non trivial as the loss $\mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j; \mathbf{W}), \hat{\mathbf{y}}_j)$ involves a min-max optimization problem (4). To this end, we consider a regularized version of $\mathrm{W}_{\mathrm{ROT}}$ by adding a negative entropy regularization term to (5). Equivalently, we consider the formulation for computing the robust OT distance between $h(\mathbf{x}_j; \mathbf{W})$ and $\hat{\mathbf{y}}_j$ as

$$\mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j; \mathbf{W}), \hat{\mathbf{y}}_j) = \min_{\gamma \in \Pi(h(\mathbf{x}_j;\mathbf{W}), \hat{\mathbf{y}}_j)} \max_{\mathbf{M} \in \mathcal{M}} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle + \lambda_\gamma \textstyle\sum_{pq} \gamma_{pq} \ln(\gamma_{pq}), \tag{17}$$

where $\lambda_\gamma > 0$ and $\mathbf{V}_\gamma = \sum_{pq} (\mathbf{l}_p - \mathbf{l}_q)(\mathbf{l}_p - \mathbf{l}_q)^\top \gamma_{pq}$ is of size $L \times L$. Here, $\mathbf{l}_p$ is the ground embedding of p-th label.

The following lemma provides an expression for the gradient of $\mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j; \mathbf{W}), \hat{\mathbf{y}}_j)$ with respect to $h(\mathbf{x}; \mathbf{W})$.

Lemma 5.1. *Let $(\gamma^*, \mathbf{M}^*(\gamma^*))$ denote the optimal solution of the robust OT problem (17). Then,*

$$\nabla_{h(\mathbf{x};\mathbf{W})} \mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j; \mathbf{W}), \hat{\mathbf{y}}_j) = \tfrac{1}{L} \mathbf{A}\mathbf{1} - \tfrac{\mathbf{1}^\top \mathbf{A} \mathbf{1}}{L^2} \mathbf{1}, \tag{18}$$

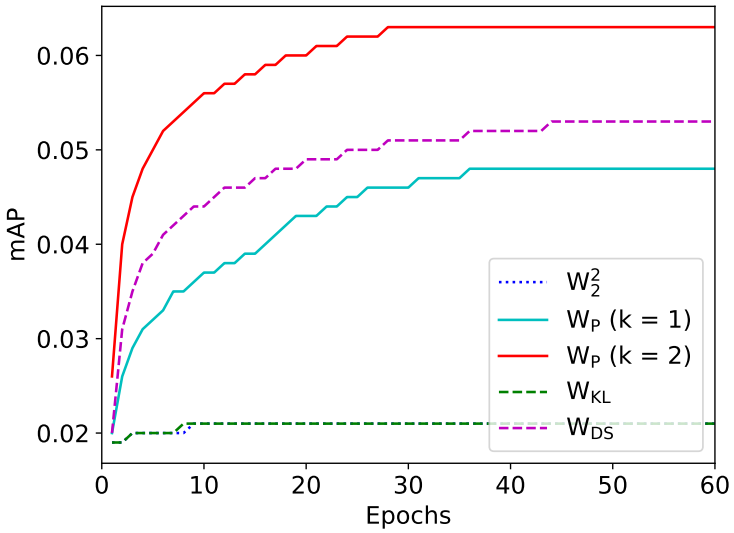

Figure 4: Evolution of mean average precision (mAP) on the Flickr tag-prediction dataset for various OT distance-based loss functions.

where $\mathbf{1}$ *is the column vector of ones of size* L *and* $\mathbf{A} = \mathbf{C}^* + \lambda_\gamma(\ln(\gamma^*) + \mathbf{1}\mathbf{1}^\top)$. *Here,* $\mathbf{C}^*_{pq} = (\mathbf{l}_p - \mathbf{l}_q)^\top \mathbf{M}^*(\gamma^*)(\mathbf{l}_p - \mathbf{l}_q)$.

The proof of Lemma 5.1 is in Section A.6.

For the multi-label setting, computation of the gradient $\nabla_{h(\mathbf{x};\mathbf{W})} \mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j;\mathbf{W}), \hat{\mathbf{y}}_j)$, shown in (18), in Lemma 5.1 costs $O(l_j L d^2)$, where l_j is the number of ground-truth labels for the j-th training instance. Using the decomposition (12), the cost reduces to $O(l_j L r^2)$. In many cases, l_j is much smaller than L.

For the multi-class setting, computation of the gradient $\nabla_{h(\mathbf{x};\mathbf{W})} \mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j;\mathbf{W}), \hat{\mathbf{y}}_j)$ costs $O(Ld^2)$ by setting $l_j = 1$. Similar to the multi-label setting, using the decomposition (12), the cost reduces to $O(Lr^2)$.

Overall, in both multi-class/multi-label settings, the cost of the gradient computation in (18) scales linearly with the number of labels L and quadratically with r. When $r \ll d$, optimization of the $\mathrm{W}_{\mathrm{ROT}}$ loss becomes computationally feasible for large-scale multi-class/multi-label instances.

6 Experiments

We evaluate the proposed robust optimal transport formulations in the supervised multi-class/multi-label setting discussed in Section 5. Our code is available at `https://github.com/satyadevntv/ROT`.

6.1 Datasets and evaluation setup

We experiment on the following three multi-class/multi-label datasets.

Animals (Lampert et al., 2009): This dataset contains 30 475 images of 50 different animals. DeCAF features (4096 dimensions) of each image are available at `https://github.com/jindongwang/`

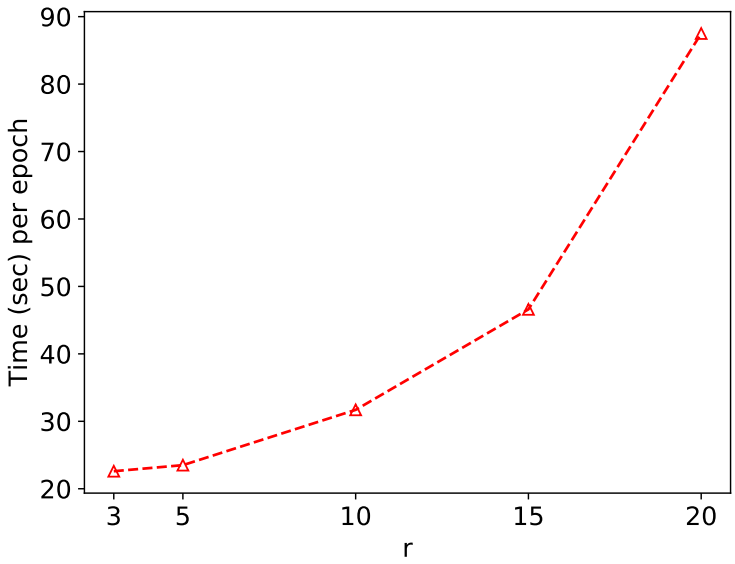

Figure 5: Time taken per epoch versus r on the Flickr dataset by our algorithm with the proposed $\mathrm{W_P}$ loss with $k = 1$.

`transferlearning/blob/master/data/dataset.md`. We randomly sample 10 samples per class for training and the rest are used for evaluation.

MNIST: The MNIST handwritten digit dataset consists of images of digits $\{0, \ldots, 9\}$. The images are of 28×28 pixels, leading to 784 features. The pixel values are normalized by dividing each dimension with 255. We randomly sample 100 images per class (digit) for training and 1000 images per class for evaluation.

Flickr (Thomee et al., 2016): The Yahoo/Flickr Creative Commons 100M dataset consists of descriptive tags for around 100M images. We follow the experimental protocol in (Frogner et al., 2015) for the tag-prediction (multi-label) problem on 1000 descriptive tags. The training and test sets consist of 10 000 randomly selected images associated with these tags. The features for images are extracted using MatConvNet (Vedaldi & Lenc, 2015). The train/test sets as well as the image features are available at `http://cbcl.mit.edu/wasserstein`.

As described in Section 5, we use the fastText word embeddings (Bojanowski et al., 2016) corresponding to the labels for computing the OT ground metric in all our evaluations.

We report the standard AUC metric for all the experiments. For the Flickr tag-prediction problem, we additionally report the mAP (mean average precision) metric. As the datasets are high dimensional, we use the low-dimensional decomposition of the Mahalanobis metric $\mathbf{M}$ (Section 3) for the proposed robust OT distance-based loss functions by randomly grouping the features. We experiment with $r \in \{5, 10, 20\}$. As baseline, we report the results with the W_2^2 distance-based loss function (2) on all the datasets. Additional experimental details and results are in Section B.

6.2 Results and discussion

We report the results of our experiments in Table 1. Overall, the robust OT loss functions provide better generalization performance than the $\mathbf{W}_2^2$ distance-based loss function. The proposed $\mathrm{W_P}$ loss

with $k = 1$ obtains the best results on all the three datasets. The proposed $\mathrm{W_P}$ loss with $k = 2$ obtains a good AUC/mAP on the Animals and the Flickr datasets but does not generalize well in MNIST. Similarly, while $\mathrm{W_{DS}}$ with $r = 5$ obtains good results, its performance decreases with r. This is because the non-linear nature of the proposed OT distance-based loss function may lead to overfitting, especially at high r.

We also report the time taken per epoch of the SGD algorithm on the Flickr dataset with $\mathrm{W_P}$ ($k = 1$) as the loss function. Our experiments are run on a machine with 32 core Intel CPU (2.1 GHz Xeon) and a single NVIDIA GeForce RTX 2080 Ti GPU (11 GB). The model computations are performed on the GPU. From Figure 5, we observe that our model scales gracefully as $O(r^2)$ as discussed in Section 4. Empirically, we show that we can obtain a good generalization performance and computational efficiency with $r \ll d$.

7 Conclusion

We have discussed robust optimal transport problems arising from Mahalanobis parameterized cost functions. A particular focus was on discussing novel formulations that are less computationally heavy from an optimization viewpoint. We also proposed a low-dimensional decomposition of the Mahalanobis metric for efficient computation of robust OT distances for high-dimensional data. An immediate result was that it allowed the use the robust OT formulations in high-dimensional multi-class/multi-label problems. Our experiments on real-world datasets suggest good performance of the robust OT loss in learning problems.

Acknowledgment

We thank J. Saketha Nath for insightful discussions on the topic.

References

Alvarez-Melis, D. and Jaakkola, T. Gromov-Wasserstein alignment of word embedding spaces. In *EMNLP*, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *ICML*, 2017.

Arora, R., Gupta, M., Kapila, A., and Fazel, M. Clustering by left-stochastic matrix factorization. In *ICML*, 2011.

Bertsekas, D. P. *Nonlinear Programming.* Athena Scientific, 1995.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. In *NeurIPS*, 2017.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. Max-sliced Wasserstein distance and its use for gans. In *CVPR*, 2019.

Dhouib, S., Redko, I., Kerdoncuff, T., Emonet, R., and Sebban, M. A Swiss army knife for minimax optimal transport. In *ICML*, 2020.

Douik, A. and Hassibi, B. Low-rank Riemannian optimization on positive semidefinite stochastic matrices with applications to graph clustering. In *ICML*, 2018.

Douik, A. and Hassibi, B. Manifold optimization over the set of doubly stochastic matrices: A second-order geometry. *IEEE Transactions on Signal Processing*, 67(22):5761–5774, 2019.

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. Learning with a wasserstein loss. In *NeurIPS*, 2015.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *AISTATS*, 2018.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

Janati, H., Cuturi, M., and Gramfort, A. Wasserstein regularization for sparse multi-task regression. In *AISTATS*, 2017.

Jawanpuria, P., Lapin, M., Hein, M., and Schiele, B. Efficient output kernel learning for multiple tasks. In *NeurIPS*, 2015.

Kantorovich, L. On the translocation of masses. *Doklady of the Academy of Sciences of the USSR*, 37:199–201, 1942.

Knight, P. A. The sinkhorn-knopp algorithm: Convergence and applications. *SIAM J. Matrix Anal. Appl.*, 30(1):261—275, 2008.

Kolouri, S., Nadjahi, K., Şimşekli, U., Badeau, R., and Rohde, G. K. Generalized sliced Wasserstein distances. In *NeurIPS*, 2019.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013.

Paty, F.-P. and Cuturi, M. Subspace robust Wasserstein distances. In *ICML*, 2019.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Qi, G.-J., Tang, J., Zha, Z.-J., Chua, T.-S., and Zhang, H.-J. An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization. In *ICML*, 2009.

Rosales, R. and Fung, G. Learning sparse metrics via linear programming. In *KDD*, 2006.

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

Sion, M. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of ACM*, 59(2):64–73, 2016.

Vedaldi, A. and Lenc, K. Matconvnet: Convolutional neural networks for matlab. In *ACM International Conference on Multimedia*, pp. 689–692, 2015.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Verlag, 2009.

Wang, X., Nie, F., and Huang, H. Structured doubly stochastic matrix for graph based clustering. In *SIGKDD*, 2016.

Zass, R. and Shashua, A. Doubly stochastic normalization for spectral clustering. In *NeurIPS*, 2006.

A Proofs

A.1 Proof of Theorem 2.1

The following result was proved in the context of output kernels for multitask learning problems by Jawanpuria et al. (2015, Theorem 4). That is,

$$\max_{\mathbf{M}\succeq\mathbf{0}} \ \langle \mathbf{V}_\gamma, \mathbf{M}\rangle - \|\mathbf{M}\|_p^p = 0.5\left(\frac{2k-1}{2k}\right)^{2k-1} \|\mathbf{V}_\gamma\|_{2k}^{2k}\,, \tag{19}$$

where $k \in \mathbb{N}$, $p = \frac{2k}{2k-1}$, and the optimal solution of (19) is $\mathbf{M}^* = \left(\frac{2k-1}{2k}\right)^{2k-1} (\mathbf{V}_\gamma)^{\circ(2k-1)}$. In the following, we use a similar proof technique for our problem with the constraint $\|\mathbf{M}\|_p \leq 1$, i.e.,

$$\mathrm{W_P}(\mu_1, \mu_2) \coloneqq \min_{\gamma\in\Pi(\mu_1,\mu_2)} f(\gamma), \tag{20}$$

where

$$f(\gamma) = \max_{\mathbf{M}\succeq\mathbf{0}, \|\mathbf{M}\|_p\leq 1} \ \langle \mathbf{V}_\gamma, \mathbf{M}\rangle\,. \tag{21}$$

Let us consider the variant of (21) without the symmetric positive semi-definite constraint, i.e.,

$$\mathbf{M}^*(\gamma) = \arg\max_{\|\mathbf{M}\|_p\leq 1} \ \langle \mathbf{V}_\gamma, \mathbf{M}\rangle \tag{22}$$

From the Hölder's inequality (for vectors), the optimal solution of (22) is $\mathbf{M}^*(\gamma) = \|\mathbf{V}_\gamma\|_q^{-q/p} (\mathbf{V}_\gamma)^{\circ(q/p)}$, where q-norm is the dual of p-norm, i.e., $1/p + 1/q = 1$.

For $k \in \mathbb{N}$ and $p = \frac{2k}{2k-1}$, we have $q = \frac{1}{2k}$. Therefore, $\mathbf{M}^*(\gamma) = \|\mathbf{V}_\gamma\|_{2k}^{1-2k} (\mathbf{V}_\gamma)^{\circ(2k-1)}$. It should be noted that the $\mathbf{M}^*(\gamma)$ for such p-norm is a symmetric positive semi-definite matrix (via application of the Schur product theorem) as $\mathbf{V}_\gamma$ is a symmetric positive semi-definite matrix. Consequently, $\mathbf{M}^*(\gamma) = \|\mathbf{V}_\gamma\|_{2k}^{1-2k} (\mathbf{V}_\gamma)^{\circ(2k-1)}$ is also the optimal solution of (21). Substituting this value of $\mathbf{M}^*(\gamma)$ in (21) leads to (6), thereby proving Theorem 2.1.

A.2 Proof of Corollary 2.2

From Theorem 2.1, we have the following result:

$$\mathrm{W_P}(\mu_1, \mu_2) = \min_{\gamma\in\Pi(\mu_1,\mu_2)} \|\mathbf{V}_\gamma\|_{2k} = \min_{\gamma\in\Pi(\mu_1,\mu_2)} \max_{\mathbf{M}\in\mathcal{M}} \ \langle \mathbf{V}_\gamma, \mathbf{M}\rangle = \max_{\mathbf{M}\in\mathcal{M}} \min_{\gamma\in\Pi(\mu_1,\mu_2)} \ \langle \mathbf{V}_\gamma, \mathbf{M}\rangle\,, \tag{23}$$

where $k \in \mathbb{N}$, $p = \frac{2k}{2k-1}$, and $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0} \text{ and } \|\mathbf{M}\|_p \leq 1\}$. Let us consider a feasible $\mathbf{M} = \frac{1}{d^{\frac{1}{p}}}\mathbf{I} \in \mathcal{M}$. Then, we have

$$\mathrm{W_P}(\mu_1, \mu_2) = \max_{\mathbf{M}\in\mathcal{M}} \min_{\gamma\in\Pi(\mu_1,\mu_2)} \langle \mathbf{V}_\gamma, \mathbf{M}\rangle \geq \min_{\gamma\in\Pi(\mu_1,\mu_2)} \left\langle \mathbf{V}_\gamma, \frac{1}{d^{\frac{1}{p}}}\mathbf{I}\right\rangle = \frac{1}{d^{\frac{1}{p}}}\mathrm{W}_2^2(\mu_1, \mu_2),$$

which gives us the lower bound result in Corollary 2.2.

The upper bound result in Corollary 2.2 may be proved as follows. We first note a result proved in (Dhouib et al., 2020):

$$\mathrm{W}_2^2(\mu_1, \mu_2) = \min_{\gamma\in\Pi(\mu_1,\mu_2)} \|\mathbf{V}_\gamma\|_{*1}\,, \tag{24}$$

where $\|\cdot\|_{*q}$ denotes the Schatten q-norm regularizer (discussed in Section 1). We also have the following result from the inequality relationships between the norms:

$$\|\mathbf{V}_\gamma\|_{*1} \geq \|\mathbf{V}_\gamma\|_{*2} = \|\mathbf{V}_\gamma\|_2 \geq \|\mathbf{V}_\gamma\|_{2k}\,. \tag{25}$$

Let γ_1 be the optimal solution of (24). Then, we have the following result from (25):

$$\mathrm{W}_2^2(\mu_1,\mu_2) = \|\mathbf{V}_{\gamma_1}\|_{*1} \geq \|\mathbf{V}_{\gamma_1}\|_{*2} = \|\mathbf{V}_{\gamma_1}\|_2 \geq \|\mathbf{V}_{\gamma_1}\|_{2k} \geq \min_{\gamma\in\Pi(\mu_1,\mu_2)} \|\mathbf{V}_\gamma\|_{2k} = \mathrm{W_P}(\mu_1,\mu_2).$$

Thus, we have proved the upper bound result in Corollary 2.2: $\mathrm{W}_2^2(\mu_1,\mu_2) \geq \min_{\gamma\in\Pi(\mu_1,\mu_2)} \|\mathbf{V}_\gamma\|_{2k} = \mathrm{W_P}(\mu_1,\mu_2)$.

A.3 Proof of Theorem 2.3

We consider the following unconstrained variant of Problem (8) and consider its optimal solutions

$$\mathbf{M}^*(\gamma) = \arg\max_{\mathbf{M}}\ \langle \mathbf{V}_\gamma, \mathbf{M}\rangle - \lambda_{\mathbf{M}}\mathrm{D_{KL}}(\mathbf{M},\mathbf{M}_0) = \mathbf{M}_0 \odot e^{\circ(\mathbf{V}_\gamma/\lambda_{\mathbf{M}})}. \tag{26}$$

The optimal solution of the above unconstrained concave maximization problem is obtained simply as the first-order critical point (by setting gradient of the objective function to zero).

It should be noted that $\mathbf{M}^*(\gamma)$ in (26) is a symmetric positive semi-definite matrix. This is because (a) $\mathbf{M}_0$ is a given symmetric positive semi-definite matrix, (b) $e^{\circ(\mathbf{V}_\gamma/\lambda_{\mathbf{M}})}$ is a symmetric positive semi-definite matrix as $\mathbf{V}_\gamma$ is symmetric and positive semi-definite, and (c) the Hadamard product of two symmetric positive semi-definite matrices is a symmetric positive semi-definite matrix (Schur product theorem). Hence, $\mathbf{M}^*(\gamma)$ in (26) is also the optimal solution of (8).

Substituting the solution (26) in Problem (4), where $f(\gamma)$ is defined as in (8), gives us the expression of $\mathrm{W_{KL}}(\mu_1,\mu_2)$ in (9). This completes the proof of Theorem 2.3.

A.4 Proof of Theorem 2.4

Consider a variant of Problem (11), i.e.,

$$\begin{aligned}\mathbf{M}^*(\gamma) &= \arg\max_{\mathbf{M}\geq\mathbf{0},\mathbf{M}\mathbf{1}=\mathbf{1},\mathbf{M}^\top\mathbf{1}=\mathbf{1}} \langle \mathbf{V}_\gamma, \mathbf{M}\rangle - \lambda_{\mathbf{M}}\mathrm{D_{KL}}(\mathbf{M},\mathbf{M}_0)\\ &= \arg\max_{\mathbf{M}\geq\mathbf{0},\mathbf{M}\mathbf{1}=\mathbf{1},\mathbf{M}^\top\mathbf{1}=\mathbf{1}} \langle \mathbf{V}_\gamma + \ln(\mathbf{M}_0), \mathbf{M}\rangle - \lambda_{\mathbf{M}}\langle \ln(\mathbf{M}), \mathbf{M}\rangle\,,\end{aligned} \tag{27}$$

where $\ln(\mathbf{M}_0)$ and $\ln(\mathbf{M})$ denotes the application of the natural logarithm function on each entry of the matrices $\mathbf{M}_0$ and $\mathbf{M}$, respectively. The difference between Problems (11) and (27) is the search space, i.e., Problem (27) does not have the symmetric positive semi-definiteness constraint on the feasible points. Thus, the search space of Problem (27) is a super-set of the search space of Problem (11). In the following, we prove that the optimal solution $\mathbf{M}^*(\gamma)$ in (27) satisfies symmetric positive semi-definiteness constraint. Thus, $\mathbf{M}^*(\gamma)$ in (27) also becomes the optimal solution of Problem (11).

It should be noted that Problem (27) has the same form as the entropic-regularized optimal transport problem studied in (Cuturi, 2013), with a symmetric cost matrix as $-(\mathbf{V}_\gamma + \ln(\mathbf{M}_0))$. As discussed in (Cuturi, 2013), the solution of the the entropic-regularized optimal transport problem

can be efficiently obtained using the Sinkhorn-Knopp algorithm (Knight, 2008). In case the cost matrix is symmetric, the optimal solution of (27) via the Sinkhorn-Knopp algorithm has the following form:

$$\mathbf{M}^*(\gamma) = \mathbf{D}\left(\mathbf{M}_0 \odot e^{\circ(\mathbf{V}_\gamma/\lambda_\mathbf{M})}\right)\mathbf{D}, \tag{28}$$

where $\mathbf{D}$ is a diagonal matrix with positive entries (Cuturi, 2013, Lemma 2 proof). Please refer to (Knight, 2008, Section 3) for a general analysis of the Sinkhorn-Knopp algorithm with symmetric matrices.

It should be noted that the optimal solution $\mathbf{M}^*(\gamma)$ of (27), whose expression is given in (28), is a symmetric positive semi-definite matrix as $\mathbf{M}_0 \odot e^{\circ(\mathbf{V}_\gamma/\lambda_\mathbf{M})}$ is a symmetric positive semi-definite matrix (see the proof of Theorem 2.3 in Section A.3).

A.5 Proof of Lemma 4.1 on gradient computation of $\nabla_\gamma f$

The proof follows directly from the Danskin's theorem (Bertsekas, 1995) and exploits the structure of $\mathbf{V}_\gamma$.

A.6 Proof of Lemma 5.1 on gradient computation for $\mathrm{W}_{\mathrm{ROT}}$

Rewriting (17) for computing the robust OT distance between $h(\mathbf{x}_j;\mathbf{W})$ and $\hat{\mathbf{y}}_j$ as

$$\begin{array}{rcl} \mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j;\mathbf{W}),\hat{\mathbf{y}}_j) & = & \min\limits_{\gamma\in\Pi(h(\mathbf{x}_j;\mathbf{W}),\hat{\mathbf{y}}_j)} \max\limits_{\mathbf{M}\in\mathcal{M}} \langle \mathbf{V}_\gamma, \mathbf{M}\rangle + \lambda_\gamma \sum_{pq}\gamma_{pq}\ln(\gamma_{pq}) \\ & = & \min\limits_{\gamma\in\Pi(h(\mathbf{x}_j;\mathbf{W}),\hat{\mathbf{y}}_j)} f(\gamma) + \lambda_\gamma\langle\gamma,\ln(\gamma)\rangle, \end{array} \tag{29}$$

where $\lambda_\gamma > 0$ and $\mathbf{V}_\gamma = \sum_{pq}(\mathbf{l}_p - \mathbf{l}_q)(\mathbf{l}_p - \mathbf{l}_q)^\top \gamma_{pq}$ is of size $L \times L$. Here, $\mathbf{l}_p$ is the ground embedding of p-th label.

Denoting $\mu_1 = h(\mathbf{x}_j;\mathbf{W})$ and $\mu_2 = \hat{\mathbf{y}}_j$, we have

$$\begin{array}{l} \mathrm{W}_{\mathrm{ROT}}(h(\mathbf{x}_j;\mathbf{W}),\hat{\mathbf{y}}_j) \\ = \min\limits_{\gamma\in\Pi(\mu_1,\mu_2)} f(\gamma) + \lambda_\gamma\langle\gamma,\ln(\gamma)\rangle \\ = \max\limits_{\nu_1,\nu_2\in\mathbb{R}^L,\Delta\in\mathbb{R}^{L\times L}\geq 0} \min\limits_{\gamma\in\mathbb{R}^{L\times L}} f(\gamma) + \lambda_\gamma\langle\gamma,\ln(\gamma)\rangle - \langle\nu_1,\gamma\mathbf{1}-\mu_1\rangle - \langle\nu_2,\gamma^\top\mathbf{1}-\mu_2\rangle - \langle\Delta,\gamma\rangle, \end{array} \tag{30}$$

where $\mathbf{1}$ is the column vector of ones of size L, $\langle\cdot,\cdot\rangle$ is the standard inner product between matrices, ν_1, ν_2, and Δ are the dual variables, and f is a non-linear convex function obtained as $f(\gamma) = \max_{\mathbf{M}\in\mathcal{M}}\langle\mathbf{V}_\gamma,\mathbf{M}\rangle$. The last equality in (30) comes from strong duality.

The problem of interest is to compute the gradient $\nabla_{\mu_1}\mathrm{W}_{\mathrm{ROT}}(\mu_1,\mu_2)$ of $\mathrm{W}_{\mathrm{ROT}}(\mu_1,\mu_2)$ with respect to μ_1. Given the optimal solution $(\nu_1^*,\nu_2^*,\Delta^*,\gamma^*,\mathbf{M}^*(\gamma^*))$, the gradient has the expression

$$\nabla_{\mu_1}\mathrm{W}_{\mathrm{ROT}}(\mu_1,\mu_2) = \nu_1^* - \frac{\langle\nu_1^*,\mathbf{1}\rangle}{L}\mathbf{1}, \tag{31}$$

where ν_1 is the partial derivative of (30) with respect to μ_1 and the second term $\frac{\langle\nu_1^*,\mathbf{1}\rangle}{L}\mathbf{1}$ is the normal component of ν_1 to the simplex set $\mathbf{1}^\top\mu_1 = 1$. Overall, $\nabla_{\mu_1}\mathrm{W}_{\mathrm{ROT}}(\mu_1,\mu_2)$ belongs to the tangent plane of the simplex $\mathbf{1}^\top\mu_1 = 1$ at μ_1.

To compute the expression for the right hand side of (31), we look at the KKT condition of (30), i.e.,

$$\nabla_\gamma f(\gamma^*) + \lambda_\gamma(\ln(\gamma^*) + \mathbf{1}\mathbf{1}^\top) - \nu_1^*\mathbf{1}^\top - \mathbf{1}\nu_2^{*\top} - \Delta^* = \mathbf{0}. \tag{32}$$

The complementary slackness condition leads to $\Delta^* \odot \gamma^* = \mathbf{0}$. We note that the entropy regularization term ensures that $\gamma^* > 0$ implying that $\Delta^* = \mathbf{0}$. Consequently, (32) boils down to

$$\begin{aligned} & \nabla_\gamma f(\gamma^*) + \lambda_\gamma(\ln(\gamma^*) + \mathbf{1}\mathbf{1}^\top) - \nu_1^*\mathbf{1}^\top - \mathbf{1}\nu_2^{*\top} = \mathbf{0} \\ \Rightarrow \quad & \nu_1^*\mathbf{1}^\top + \mathbf{1}\nu_2^{*\top} = \nabla_\gamma f(\gamma^*) + \lambda_\gamma(\ln(\gamma^*) + \mathbf{1}\mathbf{1}^\top) \\ \Rightarrow \quad & \nu_1^*\mathbf{1}^\top + \mathbf{1}\nu_2^{*\top} = \mathbf{A}, \end{aligned} \tag{33}$$

where $\mathbf{A} = \nabla_\gamma f(\gamma^*) + \lambda_\gamma(\ln(\gamma^*) + \mathbf{1}\mathbf{1}^\top) = \mathbf{C}^* + \lambda_\gamma(\ln(\gamma^*) + \mathbf{1}\mathbf{1}^\top)$. Here, $\mathbf{C}^*_{pq} = (\mathbf{l}_p - \mathbf{l}_q)^\top \mathbf{M}^*(\gamma^*)(\mathbf{l}_p - \mathbf{l}_q)$ and $\mathbf{l}_p$ is the ground embedding for p-th label. From (33), ν_1^* and ν_2^* are translation invariant, i.e., $\nu_1^* + \alpha\mathbf{1}$ and $\nu_2^* - \alpha\mathbf{1}$ are solutions for all $\alpha \in \mathbb{R}$. However, we are interested not in ν_1^*, but in $\nu_1^* - \frac{\langle \nu_1^*, \mathbf{1}\rangle}{L}\mathbf{1}$, which is unique (as its mean is $\mathbf{0}$). Below, we compute the term $\nu_1^* - \frac{\langle \nu_1^*, \mathbf{1}\rangle}{L}\mathbf{1}$ directly.

In (33), we eliminate ν_2^* by pre-multiplying with $\mathbf{1}^\top$ to obtain

$$\nu_2^{*\top} = \tfrac{1}{L}(\mathbf{1}^\top\mathbf{A} - \mathbf{1}^\top\nu_1^*\mathbf{1}^\top). \tag{34}$$

Plugging (34) in (33), we obtain

$$\nu_1^*\mathbf{1}^\top - \tfrac{1}{L}\mathbf{1}\mathbf{1}^\top\nu_1^*\mathbf{1}^\top = \mathbf{A} - \tfrac{1}{L}\mathbf{1}\mathbf{1}^\top\mathbf{A}. \tag{35}$$

Post-multiplying $\mathbf{1}$ in (35), we obtain

$$\nu_1^* - \frac{\langle \nu_1^*, \mathbf{1}\rangle}{L}\mathbf{1} = \frac{1}{L}\mathbf{A}\mathbf{1} - \frac{\mathbf{1}^\top\mathbf{A}\mathbf{1}}{L^2}\mathbf{1}.$$

This completes the proof of Lemma 5.1.

B Experiments

B.1 Multi-class/multi-label experiments

Experimental setup: All the multi-class/label learning experiments are performed in a standard setting, where the fastText embeddings are unit normalized (via 2-norm), the Sinkhorn algorithm is run for 10 iterations, the FW algorithm is run for 1 iteration (our initial experiments showed that a single FW iteration resulted in a good quality convergence), and λ_γ is 0.02 in (17), and λ_β (LMO step in Algorithm 1) is 0.2. Following (Frogner et al., 2015), we regularize the softmax model parameters by $0.0005\|\mathbf{W}\|_2^2$ in Problem (17).

Standard deviation corresponding to Table 1: Table 2 shows the mean AUC and the corresponding standard deviation obtained across five random train/test splits for the Animals and MNIST datasets.

Results on randomized grouping of features: We next evaluate the robustness of proposed OT distance-based loss functions with respect to randomized grouping of the features into r groups. As discussed in Section 3, the low-dimensional decomposition of the Mahalanobis metric $\mathbf{M}$ requires dividing the d features into r groups. In our first set of experiments, Tables 1 and 2, we randomly create r groups of the 300-dimensional fastText features. This grouping is kept consistent for all the five train-test splits of the Animals and MNIST datasets.

Table 2: Average performance across five random train/test splits on Animals and MNIST datasets.

Method	r	Animals (AUC)	MNIST (AUC)
W_2^2 loss (Frogner et al., 2015)	–	0.794 ± 0.003	0.848 ± 0.002
W_P loss with $k = 1$ (Eq. 6)	5	0.801 ± 0.002	$\mathbf{0.959} \pm 0.003$
	10	0.907 ± 0.001	0.940 ± 0.002
	20	$\mathbf{0.908} \pm 0.002$	0.922 ± 0.003
W_P loss with $k = 2$ (Eq. 6)	5	0.890 ± 0.001	0.804 ± 0.010
	10	0.878 ± 0.001	0.702 ± 0.003
	20	0.874 ± 0.001	0.603 ± 0.003
W_{KL} loss (Eq. 9)	5	0.794 ± 0.003	0.867 ± 0.034
	10	0.749 ± 0.002	0.870 ± 0.036
	20	0.794 ± 0.003	0.779 ± 0.002
W_{DS} loss (Eq. 11)	5	$\mathbf{0.908} \pm 0.002$	0.927 ± 0.002
	10	0.878 ± 0.001	0.873 ± 0.003
	20	0.845 ± 0.003	0.813 ± 0.003

In the next experiment, we fix the train-test split (the first among the five random train-test splits of the Animals and MNIST datasets) and obtain results on four additional random groupings of the features. Table 3 reports the mean AUC and the corresponding standard deviation obtained across five random groupings of features on the same train-test split. We observe that the results are similar to those reported in Table 2, thereby signifying the robustness of the proposed OT distance-based loss functions with regards to groupings of features.

B.2 Movies dataset

We also comparatively study the subspace Robust Wasserstein (SRW) distance (Paty & Cuturi, 2019) and the W_P distance (with $r = 10, k = 1$) between the scripts of seven movies. We follow the experimental protocol described in (Paty & Cuturi, 2019). The marginals are the histograms computed from the word frequencies in the movie scripts and each word is represented as a 300-dimensional fastText embedding (Bojanowski et al., 2016).

It should be noted that the range/spread of SRW (3) and W_P (6) distances are different for the same movie. Hence, Table 4 reports the normalized SRW and the normalized W_P distances for all pairs of movies. The normalization is done column-wise as follows: we divide all the SRW distances in a column by the maximum SRW distance in that column (and similarly normalize the W_P distances as well). This normalization ensures that for a given movie (representing the column), the minimum relative distance is 0 (with itself) while the maximum relative distance is 1 for both SRW and W_P.

We observe that both SRW and W_P are usually consistent in selecting the closest movie (i.e., the row corresponding with minimum non-zero distance in a column). However, W_P tends to have

Table 3: Average performance on five random groupings of features on Animals and MNIST datasets

Method	r	Animals (AUC)	MNIST (AUC)
W_P loss with $k = 1$ (Eq. 6)	5	0.805 ± 0.003	$\mathbf{0.957} \pm 0.001$
	10	0.909 ± 0.002	0.939 ± 0.001
	20	$\mathbf{0.911} \pm 0.002$	0.920 ± 0.002
W_P loss with $k = 2$ (Eq. 6)	5	0.892 ± 0.001	0.787 ± 0.010
	10	0.880 ± 0.001	0.700 ± 0.002
	20	0.875 ± 0.002	0.601 ± 0.001
W_{KL} loss (Eq. 9)	5	0.799 ± 0.003	0.806 ± 0.034
	10	0.799 ± 0.003	0.801 ± 0.039
	20	0.799 ± 0.003	0.781 ± 0.001
W_{DS} loss (Eq. 11)	5	$\mathbf{0.911} \pm 0.002$	0.928 ± 0.002
	10	0.884 ± 0.003	0.868 ± 0.010
	20	0.855 ± 0.006	0.817 ± 0.013

a wider spread of distances, i.e., the difference in the distances corresponding to the closest and the furthest movies. As an example, SRW computes similar distances for the pairs (Kill Bill Vol.1, Kill Bill Vol.2) and (Inception, The Martian) while W_P gives the (Kill Bill Vol.1, Kill Bill Vol.2) pair a much lower relative distance (which seems more reasonable as they are sequels).

Table 4: Normalized distances between movie scripts. Each cell indicates the SRW_2^2/W_P distances respectively. Here, D = Dunkirk, G = Gravity, I = Interstellar, KB1 = Kill Bill Vol.1, KB2 = Kill Bill Vol.2, TM = The Martian, and T = Titanic.

	D	G	I	KB1	KB2	TM	T
D	0.000/0.000	0.906/0.943	0.911/0.951	0.995/1.000	0.995/0.998	0.964/1.000	**0.924**/0.931
G	0.911/0.931	0.000/0.000	0.847/0.880	1.000/1.000	1.000/1.000	0.907/0.858	1.000/0.978
I	0.916/0.901	**0.847**/**0.846**	0.000/0.000	0.995/0.953	1.000/0.961	**0.876**/**0.814**	0.978/**0.898**
KB1	0.965/0.972	0.966/0.985	0.961/0.978	0.000/0.000	**0.808**/**0.673**	0.984/0.964	0.973/0.945
KB2	1.000/0.984	1.000/1.000	1.000/1.000	**0.837**/**0.683**	0.000/0.000	1.000/0.960	0.978/0.948
TM	0.921/1.000	0.862/0.870	**0.833**/0.859	0.969/0.992	0.951/0.973	0.000/0.000	0.989/1.000
T	**0.842**/**0.814**	0.906/0.867	0.887/**0.828**	0.913/0.849	0.887/0.840	0.943/0.874	0.000/0.000