OPT2020: 12th Annual Workshop on Optimization for Machine Learning

Efficient robust optimal transport: formulations and algorithms

Pratik Jawanpuria PRATIK.JAWANPURIA@MICROSOFT.COM
Microsoft, India

N T V Satya Dev SATYADEV@VAYVE.IN
Vayve Technologies, India

Bamdev Mishra BAMDEVM@MICROSOFT.COM
Microsoft, India

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 positive semi-definite Mahalanobis metric. In particular, we study several different regularizations on the Mahalanobis metric – element-wise p-norm, KL-divergence, and 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 the Frank-Wolfe algorithm. Empirical results on real-world datasets show the efficacy of our approach.

1. Introduction

Optimal transport (OT) has become a popular tool in diverse machine learning applications such as domain adaptation [5, 13], multi-task learning [12], natural language processing [1], and multi-label classification [10], to name a few. The classical discrete OT problem, also popularly known as the earth mover's distance [17], may be formulated as follows:

$$\mathrm{W}_c(\mu_1, \mu_2) = \min_{\gamma \in \Pi(\mu_1, \mu_2)} \langle \gamma, \mathbf{C} \rangle \,, \tag{1}$$

where $\mathbf{C} \in \mathbb{R}_+^{m \times n}$ is the ground cost matrix between the source distribution's samples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^m \in \mathbb{R}^{d \times m}$ and the target distribution's samples $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n \in \mathbb{R}^{d \times n}$, the (i, j)-th entry of $\mathbf{C}$ is $c(\mathbf{x}_i, \mathbf{y}_j)$, $c(\mathbf{x}, \mathbf{y})$ is the given ground cost function, μ_1 and μ_2 are the given discrete marginal distributions of the source and target distributions, respectively, and $\Pi(\mu_1, \mu_2)$ is the set of feasible joint transportation plan: $\Pi(\mu_1, \mu_2) = \{\gamma \in \mathbb{R}^{m \times n} : \gamma \geq \mathbf{0}; \gamma \mathbf{1} = \mu_1; \gamma^\top \mathbf{1} = \mu_2\}$. The special case of $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$ (squared Euclidean distance) is popularly denoted by W_2^2 (the 2-Wasserstein distance) and can be reformulated as follows [14]: $\mathrm{W}_2^2(\mu_1, \mu_2) = \min_{\gamma \in \Pi(\mu_1, \mu_2)} \langle \mathbf{V}_\gamma, \mathbf{I} \rangle$, where $\mathbf{V}_\gamma = \sum_{ij} (\mathbf{x}_i - \mathbf{y}_j)(\mathbf{x}_i - \mathbf{y}_j)^\top \gamma_{ij}$ and $\mathbf{I}$ is the identity matrix.

Recently, [14] propose a robust variant of the W_2^2 distance, termed as the Subspace Robust Wasserstein (SRW) distance, as follows: $\mathrm{SRW}_k^2(\mu_1, \mu_2) = \max_{\mathbf{M} \in \mathcal{M}} \min_{\gamma \in \Pi(\mu_1, \mu_2)} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle$, where the domain $\mathcal{M}$ is defined as $\mathcal{M} = \{\mathbf{M} : \mathbf{0} \preceq \mathbf{M} \preceq \mathbf{I}; \mathrm{trace}(\mathbf{M}) = k\}$. It should be

noted that $\langle \mathbf{V}_\gamma, \mathbf{M} \rangle = \sum_{i,j} \gamma_{ij} c_{\mathbf{M}}(\mathbf{x}_i, \mathbf{y}_j)$, where $c_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top \mathbf{M}(\mathbf{x} - \mathbf{y})$ is a Mahalanobis metric parameterized cost function. [7] also study the above form of Mahalanobis metric parameterized cost functions in the robust OT setting, but with the domain $\mathcal{M}$ defined as $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0}; \|\mathbf{M}\|_{*p} = 1\}$, where $\|\cdot\|_{*p}$ denotes the Schatten p-norm regularizer, i.e., $\|\mathbf{M}\|_{*p} := (\sum_i \sigma_i(\mathbf{M})^p)^{\frac{1}{p}}$. Here, $\sigma_i(\mathbf{M})$ denotes the i-th largest eigenvalue of $\mathbf{M}$. Both [7, 14] pose their Mahalanobis metric parameterized robust optimal transport problems as an optimization problem over the metric $\mathbf{M}$. This involves satisfying the positive semi-definite constraint, which typically requires costly eigendecomposition operation of $d \times d$ matrices in each step costing $O(d^3)$.

In this work, we study novel robust OT formulations where the cost function is parameterized by the Mahalanobis metric $\mathbf{M} \succeq \mathbf{0}$. For a class of regularizers on $\mathbf{M}$, we show that the problem may be solved by dropping the positive semi-definiteness constraint on the metric $\mathbf{M}$ as the resulting optimal solution $\mathbf{M}^*$ satisfies $\mathbf{M}^* \succeq \mathbf{0}$. This considerably simplifies our optimization methodology and brings down our the per-iteration computational cost of learning the $d \times d$ metric $\mathbf{M}$ to $O(d^2)$. The proposed class of regularizers on the Mahalanobis metric $\mathbf{M}$ include entry-wise p-norm for $p \in (1, 2]$, the KL-divergence, and the doubly stochastic constraint. It should be noted that $O(d^2)$ computations may also be impractical for high dimensional data. We, therefore, additionally propose a suitable low-dimensional decomposition of the Mahalanobis metric for the studied robust OT problems, resulting in the per-iteration computational cost of $O(r^2)$, where $r \ll d$. We view the robust OT min-max optimization problems as non-linear OT (minimization) problems and propose an efficient Frank-Wolfe algorithm for solving them. Empirical results on the Yahoo Flickr Creative Commons tag-prediction dataset illustrates the effectiveness of our approach.

A longer version of the manuscript is available at https://arxiv.org/pdf/2010.11852.pdf. Our code is available at https://github.com/satyadevntv/ROT4C.

2. Novel formulations for robust optimal transport

In this section, we propose three novel formulations of the Mahalanobis metric parameterized robust optimal transport problem, which may be rewritten as

$$\mathrm{W}_{\mathrm{ROT}}(\mu_1, \mu_2) := \min_{\gamma \in \Pi(\mu_1, \mu_2)} f(\gamma), \tag{2}$$

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

$$f(\gamma) := \max_{\mathbf{M} \in \mathcal{M}} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle . \tag{3}$$

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 (2) is a convex optimization problem. First-order methods for solving (2) requires computing the (sub-)gradient $\nabla f(\gamma)$, which can be obtained in terms of an optimal solution $\mathbf{M}^*(\gamma)$ of (3) by using the Danskin's theorem [3]. Since problem (3) involves the positive semi-definite constraint, computing $\mathbf{M}^*(\gamma)$ usually involves costly eigendecomposition (or equivalent operations) of $d \times d$ matrices.

We show that for the proposed family of regularizers $\Omega(\cdot)$, one can drop the positive semi-definite constraint in problem (3) as the optimal solution $\mathbf{M}^*$ of the resulting problem automatically satisfies $\mathbf{M}^* \succeq \mathbf{0}$. This considerably simplifies our optimization formulations.

2.1. Element-wise p-norm regularization on $\mathbf{M}$

We consider $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0} \text{ and } \|\mathbf{M}\|_p \leq 1\}$ in (3), where $\|\mathbf{M}\|_p = (\sum_{ij} |\mathbf{M}_{ij}|^p)^{\frac{1}{p}}$ denotes the element-wise p-norm on the matrix $\mathbf{M}$. As $\mathbf{V}_\gamma$ is the second-order moment matrix of the displacements (associated with a transport plan), 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 results from $\mathbf{M} = \mathbf{I}$, i.e., it 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 for $p = 2$. For $p \in [1, 2)$, 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 [15, 16]. The following result provides an efficient reformulation of the robust OT problem (2) 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 $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0} \text{ and } \|\mathbf{M}\|_p \leq 1\}$. Consider the following optimization problem:*

$$\mathrm{W}_{\mathrm{E}}(\mu_1, \mu_2) := \min_{\gamma \in \Pi(\mu_1, \mu_2)} \|\mathbf{V}_\gamma\|_{2k}, \tag{4}$$

where $\mathbf{V}_\gamma = \sum_{ij} (\mathbf{x}_i - \mathbf{y}_j)(\mathbf{x}_i - \mathbf{y}_j)^\top \gamma_{ij}$. Then, Problem (4) is equivalent to Problem (2) and the objectives of (2) and (4) are equal for any feasible γ. For a given $\gamma \in \Pi(\mu_1, \mu_2)$, the optimal solution of (3) 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., $\mathbf{A}^{\circ(k)}(s,t) = \mathbf{A}(s,t)^k$.

Computing optimization ingredients efficiently: An optimal solution γ^* of Problem (4) is also an optimal solution of the following problem: $\mathrm{W}_{\mathrm{E}}^{2k} = \min_{\gamma \in \Pi(\mu_1, \mu_2)} \|\mathbf{V}_\gamma\|_{2k}^{2k}$. From an optimization perspective, the gradient computation for $\mathrm{W}_{\mathrm{E}}^{2k}$ is simpler than that of Problem (4).

2.2. Doubly-stochastic regularization on $\mathbf{M}$

Positive semi-definite matrices having each row (and consequently each column) lie on the simplex are also known as positive semi-definite stochastic matrices in the literature. Applications such as graph clustering and community detection applications involve learning such matrices [2, 8, 9, 20, 21].

We consider the robust OT problem (2) in which the metric $\mathbf{M}$ is a positive semi-definite stochastic matrix, i.e.,

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

where $\mathcal{M} = \{\mathbf{M} : \mathbf{M} \succeq \mathbf{0}; \mathbf{M} > \mathbf{0}; \mathbf{M}\mathbf{1} = \mathbf{1}\}$. It should be noted that optimization over the set of positive semi-definite stochastic matrices is non-trivial and computationally challenging. We show, however, that Problem (5) can be solved efficiently by adding a negative entropy regularization term. To this end, we propose to solve for the following $\tilde{f}(\gamma)$ instead:

$$\tilde{f}(\gamma) = \arg\max_{\mathbf{M} \in \mathcal{M}} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle - \lambda \sum_{ij} \mathbf{M}_{ij} \ln \mathbf{M}_{ij}, \tag{6}$$

where $\lambda > 0$ is a small regularization parameter. Our next result discusses the solution of (6).

Theorem 2.2 *For a given $\gamma \in \Pi(\mu_1, \mu_2)$, the optimal solution of (6) has the form: $\mathbf{M}^*(\gamma) = \mathbf{D}\left(e^{\circ(\mathbf{V}_\gamma/\lambda)}\right)\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with positive entries.*

The matrix $\mathbf{D}$ in Theorem 2.2 may be efficiently computed using the Sinkhorn algorithm [6].

3. Low-dimensional decomposition of $\mathbf{M}$

In Section 2, we discuss several regularization on the Mahalanobis metric $\mathbf{M}$ that lead to efficient computation of $\mathbf{M}^*(\gamma)$, i.e., solving (3). However, given $\mathbf{V}_\gamma$, computing $\mathbf{M}^*(\gamma)$ with such regularization still requires $O(d^2)$ computations, which though linear in the size of $\mathbf{M}$, may be prohibitive for high-dimensional data. To alleviate such concerns, we propose to decompose the Mahalanobis metric $\mathbf{M}$ as follows:

$$\mathbf{M} = \mathbf{B} \otimes \mathbf{I}_{d_1}, \tag{7}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{I}_{d_1}$ denotes the identity matrix of size d_1, and $\mathbf{B} \succeq \mathbf{0}$ is a $r \times r$ positive semi-definite matrix such that $d = d_1 r$, where $d_1 \gg r$. The proposed decomposition induces the following interesting reformulation of the objective in (3): $\langle \mathbf{V}_\gamma, \mathbf{B} \otimes \mathbf{I}_{d_1} \rangle = \langle \sum_{ij} \gamma_{ij} (\mathbf{x}_i - \mathbf{y}_j)(\mathbf{x}_i - \mathbf{y}_j)^\top, \mathbf{B} \otimes \mathbf{I}_{d_1} \rangle = \langle \mathbf{U}_\gamma, \mathbf{B} \rangle$, where $\mathbf{U}_\gamma = \sum_{ij} \gamma_{ij} (\mathbf{X}_i - \mathbf{Y}_j)^\top (\mathbf{X}_i - \mathbf{Y}_j)$, and $\mathbf{X}_i$ and $\mathbf{Y}_j$ are $d_1 \times r$ matrices obtained by reshaping the vectors $\mathbf{x}_i$ and $\mathbf{y}_j$, respectively.

We observe that the proposed decomposition of the Mahalanobis metric divides the d features into r groups, each with d_1 input features. Therefore, the 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 regularization on the metric $\mathbf{M}$, among the ones discussed in Section 2, transforms into an equivalent regularization on the "group metric" $\mathbf{B}$. Thus, with the proposed decomposition $\mathbf{M} = \mathbf{B} \otimes \mathbf{I}_{d_1}$, the function $f(\gamma)$ in the robust optimal transport problem (2) may be equivalently re-written as: $f(\gamma) = \max_{\mathbf{B} \in \mathcal{M}} \langle \mathbf{U}_\gamma, \mathbf{B} \rangle$, where $\mathcal{M} = \{\mathbf{B} : \mathbf{B} \succeq \mathbf{0}; \Omega(\mathbf{B}) \leq 1\}$ and $\Omega(\cdot)$ is a regularization on $\mathbf{B}$, as discussed in Section 2.

4. Proposed algorithm

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 (2) is with the Frank-Wolfe 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 [11].

The proposed algorithm for (2) is shown in Algorithm 1. The LMO step boils down solving the optimal transport problem (1), where the cost matrix $\mathbf{C}$ is replaced by $\nabla f(\gamma)$. When regularized with an entropy regularization term, the LMO admits a computationally efficient solution using the Sinkhorn iterations [6].

Compuation of $\nabla f(\gamma)$. We begin by noting that that $\mathbf{V}_\gamma = \mathbf{Z} \mathrm{Diag}(\mathrm{vec}(\gamma)) \mathbf{Z}^\top$, where $\mathbf{Z}$ is a $d \times mn$ matrix with (i,j)-th column as $(\mathbf{x}_i - \mathbf{y}_j)$, $\mathrm{Diag}(\cdot)$ acts on a vector and outputs the corresponding diagonal matrix, and $\mathrm{vec}(\cdot)$ vectorizes a matrix in the column-major order. Using the Danskin's theorem [3],the expression of the gradient $\nabla f(\gamma)$ is $\nabla f(\gamma) = \mathrm{vec}^{-1}(\mathrm{diag}(\mathbf{Z}^\top \mathbf{M}^*(\gamma) \mathbf{Z}))$. Here, $\mathrm{diag}(\cdot)$ extracts the diagonal (vector) of a square matrix and vec^{-1} reshapes a vector into a matrix and $\mathbf{M}^*(\gamma)$ is the solution to the problem $\max_{\mathbf{M} \in \mathcal{M}} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle$ for a given γ. For the scenario discussed in Section 3, a similar expression of $\nabla f(\gamma)$ is obtained when $f(\gamma) = \max_{\mathbf{B} \in \mathcal{M}} \langle \mathbf{U}_\gamma, \mathbf{B} \rangle$.

Algorithm 1 Proposed FW algorithm for (2)

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

5. Flickr tag-prediction: Learning with robust Wasserstein loss

Frogner et al. [10] propose using the Wasserstein distance as a loss function (between the ground truth and the predictions for a given instance) in the multi-label classification setting, where both the (normalized) ground truth (μ_t) and the prediction via softmax function (μ_p) lie on a $L-1$ dimensional simplex. Here, L is the number of labels. The Wasserstein loss measures the distance between μ_t and μ_p while respecting the ground cost function c, which captures the relationship between different labels. We demonstrate the effectiveness of the proposed robust Wasserstein distances (2) as a loss function in this setting.

Computing the gradient of the robust Wasserstein loss function: Learning with the proposed robust Wasserstein loss requires computation the gradient of the robust Wasserstein distance (2) with respect to the predictions μ_p, i.e., $\nabla_{\mu_p} \mathrm{W}_{\mathrm{ROT}}(\mu_p, \mu_t)$. We compute it as follows: for a given μ_t and μ_p, we solve for (2) using the proposed FW algorithm (Section 4) and obtain the optimal γ^* and the corresponding $\mathbf{M}^* = \mathbf{M}^*(\gamma^*)$. With the known optimal Mahalanobis metric $\mathbf{M}^*$, the robust OT problem (2) reduces to the classical OT problem. Hence, in this case, the expression of $\nabla_{\mu_p} \mathrm{W}_{\mathrm{ROT}}(\mu_p, \mu_t)$ is same as the gradient expression of the Wasserstein loss proposed in [10].

Experimental setup: We follow the multi-label experimental protocol of [10] on the Yahoo/Flickr Creative Commons 100M dataset [18]. The goal is to predict the tags of the given images, i.e., words describing the given images. The number of labels is 1000 and the train/test sets consist of 10 000 images each. The features for images (available at `http://cbcl.mit.edu/wasserstein`) are extracted using MatConvNet [19]. The features of the tags, over which the Mahalanobis metric is learned for the robust Wasserstein loss, is obtained from the 300-dimensional fastText embeddings [4].

Results: Table 1 reports the standard AUC obtained with the proposed robust Wasserstein loss functions. We note that the proposed robust Wasserstein distances outperform the 2-Wasserstein distance based loss function.

References

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

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

[3] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.

$r \backslash$robust-OT-loss	$\mathrm{W_E}$ with $k=1$	$\mathrm{W_E}$ with $k=2$	$\mathrm{W_{DS}}$
$r=5$	0.706	0.760	0.724
$r=10$	0.745	0.742	0.668
$r=20$	0.768	0.737	0.628

Table 1: Standard AUC obtained using the proposed robust Wasserstein loss function on the Flickr tag-prediction problem. The AUC obtained using the W_2^2 loss function (corresponding to the 2-Wasserstein distance) is 0.649.

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

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

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

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

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

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

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

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

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

[13] J. S. Nath and P. Jawanpuria. Statistical optimal transport posed as learning kernel mean embedding. Technical report, NeurIPS, 2020.

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

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

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

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

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

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

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

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