# A SURVEY ON OPTIMAL TRANSPORT FOR MACHINE LEARNING: THEORY AND APPLICATIONS

#### A PREPRINT

#### **Luis Caicedo Torres**

Department of Mathematics and Statistics Florida International University Miami, FL, 33199 lcaic005@fiu.edu

#### Luiz Manella Pereira

Knight Foundation School of Computing and Information Sciences
Florida International University, solid lab
Miami, FL,33199
lpere339@fiu.edu

#### M. Hadi Amini

Knight Foundation School of Computing and Information Sciences
Sustainability, Optimization, and Learning for InterDependent networks laboratory (solid lab)
Florida International University
Miami, FL, 33199
moamini@fiu.edu

#### **ABSTRACT**

Optimal Transport (OT) theory has seen an increasing amount of attention from the computer science community due to its potency and relevance in modeling and machine learning. It introduces means that serve as powerful ways to compare probability distributions with each other, as well as producing optimal mappings to minimize cost functions. Therefor, it has been deployed in computer vision, improving image retrieval, image interpolation, and semantic correspondence algorithms, as well as other fields such as domain adaptation, natural language processing, and variational inference. In this survey, we propose to convey the emerging promises of the optimal transport methods across various fields, as well as future directions of study for OT in machine learning. We will begin by looking at the history of optimal transport and introducing the founders of this field. We then give a brief glance into the algorithms related to OT. Then, we will follow up with a mathematical formulation and the prerequisites to understand OT, these include Kantorovich duality, entropic regularization, KL Divergence, and Wassertein barycenters. Since OT is a computationally expensive problem, we then introduce the entropy-regularized version of computing optimal mappings, which allowed OT problems to become applicable in a wide range of machine learning problems. In fact, the methods generated from OT theory are competitive with the current state-of-the-art methods. The last portion of this survey will analyze papers that focus on the application of OT within the context of machine learning. We first cover computer vision problems; these include GANs, semantic correspondence, and convolutional Wasserstein distances. Furthermore, we follow this up by breaking down research papers that focus on graph learning, neural architecture search, document representation, and domain adaptation. We close the paper with a small section on future research. Of the recommendations presented, three main problems are fundamental to allow OT to become widely applicable but rely strongly on its mathematical formulation and thus are hardest to answer. Since OT is a novel method, there is plenty of space for new research, and with more and more competitive methods (either on an accuracy level or computational speed level) being created, the future of applied optimal transport is bright as it has become pervasive in machine learning.

#### 1 Introduction

The Optimal Transport problem sits at the intersection of various fields, including probability theory, PDEs, geometry, and optimization theory. It has seen a natural progression in its theory from when Monge first posed the problem in 1781 [24]. Now, it serves as a powerful tool due to its natural formulation in various contexts. It has recently seen a wide range of applications in computer science—most notably in computer vision, but also in natural language processing and other areas. Different elements such as the Convolutional Wasserstein Distance [36] and the Minibatch Energy Distance [4] have made significant improvements on image interpolation, heat maps, and GANs. These are examples of some problems in machine learning that are being recast using Optimal Transport elements, such as Wasserstein distance being used as an error measure for comparing different probability distributions. We note the effectiveness with which optimal transport deals with both discrete and continuous problems and the easy transition between the two classes of problems. The powerful tools from convex geometry and optimization theory have made optimal transport more viable in applications. To that extent, we note the remarkable implementation of Sinkhorn's algorithm to significantly speed up computation of Wasserstein distances [10].

Although the theory is well-developed [42], much work is being made in determining the state-of-the-art algorithms for computing optimal transport plans under various conditions. In this survey, we explore the main tools from the theory and summarize some of the major advancements in its application. While it is not all-encompassing, we aim to provide an application-focused summary.

The rest of this paper is organized as follows: Section 2 provides an overview of algorithms from different applications and major breakthroughs in computation. Section 3 presents a brief history of the topic. Section 4 details some mathematical formalism. Section 5 reviews ways to overcome the computational challenges. Section 6 and on then explores applications of OT to different fields, most notably in GANs and general image processing. We then conclude with remarks and proposed directions and close with open problems.

The interested reader can dive deeper into the rich OT material using some superb books such as [42], [41], [26], [33].

## 2 OT Algorithms at a Glance

Application	Publication	Metric Employed	Year
Computations	Sinkhorn Entropy-Reg OT [10]	Ent-Reg W-Distance	2013
Computations	2-W Barycenters [11]	Ent-Reg W-Distance	2014
Comp. Vision	Conv-W Dist [36]	Conv-W	2015
Comp. Vision	WGANs [4]	EMD	2017
Comp. Vision	OT-GAN [31]	MED	2018
Graphs	GWL [18]	Gromov - W Dist	2018
Domain Adaptation	GCG [12]	Ent-Reg W-Distance	2016

Table 1: OT Algorithms in Machine Learning Presented

An Overview of the algorithms presented in detail. Abbreviations used: Entropy Regularized Wasserstein Distance (Ent-Reg W-Distance), Minibatch Energy Distance (MED), Convolutional Wasserstein Distance (Conv-W), Gromov Wasserstein Distance (Gromov-W Dist) Earth Mover Distance (EMD), Domain Adaptation (Dom. Adap.), 2-Wasserstein (2-W), Gromov-Wasserstein Learning (GWL), Generalized Conditional Gradient (GCG)

## 3 History

The central idea of Optimal Transport (OT) can be found in the work by French geometer Gaspard Monge. In his paper, *Mémoire sur la théorie des déblais et des remblais*, published in 1781, Monge asked the question: How do I move a pile of earth (some natural resource) to a target location with the least amount of effort, or cost [24]? The idea was to find a better way of optimizing such cost that was not simply iterating through every possible permutation of supplier vs. receiver and choosing the one with the lowest cost. One of the major breakthroughs following Monge's work was by Russian mathematician Leonid Vitaliyevich Kantorovich who was the founder of linear programming. His research in optimal resource allocation, which earned him his Nobel Prize in Economics, led him to study optimal coupling and duality, thereby recasting some parts of the OT problem into a linear programming problem. Kantorovich's work led to the renaming of optimal coupling between two probability measures as the Monge-Kantorovich problem.

After Kantorovich, the field of OT gained traction and its applications expanded to several fields. For example, while John Mather worked on Lagrangian dynamical systems, he developed the theory of action-minimizing stationary measures in phase space, which led to the solution of certain Monge-Kantorovich problems [22]. Although he did not make the connection between his work and OT, Buffoni and Bernard in their paper *Optimal mass transportation and Mather theory* showed the existence of an optimal transport map while studying the "Monge transportation problem when the cost is the action associated to a Lagrangian function on a compact manifold [6]."

Several other names helped expand the field OT. For example, Yann Brenier introduced optimal coupling to his research in incompressible fluid mechanics, thus linking the two fields. Mike Cullen introduced OT in meteorology while working on semi-geostrophic equations. Both Brenier's and Cullen's work brought forth the notion that there is a connection, previously not expected, between OT and PDEs. Fields medalist Cédric Villani also contributed much to the field in connection with his work in statistical mechanics and the Boltzmann equation.

Recently, OT is being applied in several fields, including Machine Learning (ML). It started with image processing by utilizing color histograms of images (or gray images) and Wasserstein's distance to compute the similarity between images. Then, it was followed by shape recognition [25, 13, 2]. For example, in *A Metric for Distributions with Applications to Image Databases*, Rubner et al. introduced a new distance between two distributions, called Earth Mover's Distance (EMD), which reflects the minimal amount of work that must be performed to transform one distribution into the other by moving "distribution mass" around [29, 30]. Next, Haker et al. introduced a method for computing elastic registration and warping maps based on the Monge-Kantorovich theory of OT [16, 17].

Due to the important role of matrix factorization in ML, it was a natural progression to use OT as the divergence component of Nonnegative Matrix Factorization (NMF) [32]. In 2014, Solomon et al., looked at the applications of OT in semi supervised learning in their paper Wasserstein Propagation for Semi-Supervised Learning [38]. Other applications have been utilizing OT in mappings between distributions; more specifically, a recent paper was published on using Wasserstein's metric in variational inference, which lies at the heart of ML [3].

More recently, researchers have made advancements in the theory of OT with Marco Cuturi proposing methods to solve approximations of the OT problems by introducing a regularization term [10]. The field is now more active than ever, with researchers extending the theories that work for low-dimensional ML problems into high-dimensional problems, bringing forth several complex theoretical and algorithmic questions [34].

## 4 Mathematical Formalism

#### 4.1 Problem Statement

Given a connected compact Riemannian manifold M, Optimal Transport Plans (OT plans) offer a way to mathematically formulate the mapping of one probability measure  $\mu_0$  onto another probability measure  $\mu_1$ . These plans  $\pi$  are couplings that obey mass conservation laws and therefore belong to the set

$$\Pi(\mu_0, \mu_1) = \{ \pi \in \text{Prob}(M \times M) | \pi(\cdot, M) = \mu_0, \pi(M, \cdot) = \mu_1 \}$$

Here,  $\Pi$  is meant to be the set of all joint probabilities that exhibit  $\mu_0$  and  $\mu_1$  as marginal distributions. The OT plan  $\pi(x,y)$  seeks to transport mass from point x to point y. This formulation allows for mass-splitting which is to say that the optimal transport map can take portions of the mass at point x to multiple points  $y_i$ . Kantorovich sought to rephrase the Monge question into a minimization of a linear functional

$$\pi \to \inf \int_{M \times M} c(x, y) d\pi(x, y)$$
 (1)

on the nonempty and convex  $\Pi$  and appropriate cost function c. We note that some formulations accommodating multiple cost have also been proposed, e.g. [35]. Alternatively, these OT plans will minimize the distance between two measures denoted formally as the 2-Wasserstein Distance, where d is a metric:

$$W_2^d(\mu_0, \mu_1) = \inf_{\pi \in \prod(\mu_0, \mu_1)} \left( \int_{M \times M} d(x, y)^2 d\pi(x, y) \right)^{1/2}$$
 (2)

This distance defines a metric<sup>1</sup> as shown in Villani's book [42]. This distance metric will be integral to applications as we will see that it offers a new way to define loss functions. The goal is to find, or approximate, the optimal transport plan,  $\pi$ .

#### 4.2 Kantorovich Duality

Duality arguments are central to both the theoretical and numerical arguments in the OT framework. Kantorovich noticed that the minimization of the linear functional problem emits a dual problem. Here, let c denote a lower semicontinuous cost function,  $\mu_0$  and  $\mu_1$  denote marginal probabilities, and  $\Pi$  be the set of all probability measures on  $M \times M$  which emit  $\mu_0$  and  $\mu_1$  as marginals. Then, for continuous  $\phi(x)$ ,  $\psi(y)$ , we have that

$$\inf_{\Pi(\mu_0,\mu_1)} \int_{M \times M} c(x,y) d\pi(x,y) = \sup_{\phi,\psi} \int_M \phi(x) d\mu_0 + \int_M \psi(y) d\mu_1$$
 (3)

The right-hand side of the equation is known as the *dual problem* of the minimization problem and is a very useful tool in proving consequences regarding optimal transport maps. A proof of this result, along with further discussion, can be found in [42].

#### 4.3 Entropic Regularization

We can define the entropy of a coupling on  $M \times M$  by the negative energy functional coming from information theory:

$$H(\pi) = -\int \int_{M \times M} \pi(x, y) \ln(\pi(x, y)) dx dy \tag{4}$$

This entropy essentially tracks information loss of a given estimate versus the true value as it proves a lower bound for the square loss error. Then, we can consider the entropy-regularized Wasserstein distance:

$$W_{2,\gamma}^{2}(\mu_{0},\mu_{1}) = \inf_{\pi \in \Pi(\mu_{0},\mu_{1})} \left[ \int \int_{M \times M} d(x,y)^{2} d\pi(x,y) - \gamma H(\pi) \right]$$
 (5)

Cuturi proved this regularized distance offers a transport plan that is more spread out and also offers much faster computational convergence convergence [10]. This computational breakthrough will be pivotal in the tractability of Wasserstein-distance dependent algorithms.

## 4.4 KL Divergence

A lot of results for optimal transport maps can be related to the familiar KL divergence. If we define p(x) and q(x) as probability distributions given a random variable x over a manifold of distributions, then we define the KL divergence as:

$$D_{KL}(p(x)|q(x)) := \int p(x) \left( \ln \frac{p(x)}{q(x)} \right) dx \tag{6}$$

#### 4.5 Wasserstein barycenters

The barycenter problem is central to the interpolation of points in Euclidean space. Agueh and Carlier present the analog in Wasserstein space, proving its existence, uniqueness, and providing characterizations [1]. The analog is presented as the solution to the minimization of a convex combination problem

$$\inf_{\mu} \sum_{i=1}^{p} \alpha_i W_2^2(\mu_i, \mu) \tag{7}$$

where  $\mu_i$  are probability measures and the  $\alpha_i$ 's, known as barycentric coordinates, are nonnegative and sum to unity. These conclusions are derived from considering the problem dual to the problem and desirable properties of the Lengendre-Fenchel transform as well as conclusions from convex geometry. These barycenters are also uniquely characterized in relation to Brennier maps which offers direct formulation as a push forward operator. Barycenteres will play a major role in applications such as the interpolation of images under transport maps as in [36]. Computing these barycenters is discussed in the Computational Challenges section.

<sup>&</sup>lt;sup>1</sup>Here, we mean a metric in the mathematics sense, i.e. a function  $d(\cdot, \cdot): M \times M \to \mathbb{R}_+$  that is positive definite, symmetric, and subadditive on a metrizable space M. See Appendix A for more details.

## **Computational Challenges**

One the biggest challenges in the implementation of optimal transport has been its computational cost. One widely used implementation of Sinkhorn's algorithm was formulated by Cuturi significantly decreased computation cost [10]. In the following, KL denotes the Kullback-Leibler divergence, U denotes the transport polytope of transport plans P that emit r and c as marginal distributions:  $U(r,c)=\{P\in\mathbb{R}^{d\times d}_+|P\mathbb{1}_d=r,P^T\mathbb{1}_d=c\}; \text{ and } U_\alpha(r,c)=\{P\in\mathbb{R}^{d\times d}_+|P\mathbb{1}_d=r,P^T\mathbb{1}_d=c\}; \text{ and } U_\alpha(r,c)=r\}; \text{ and } U_\alpha(r,c)=r\}; \text{ and } U_$  $U(r,c)|KL(P|rc^T) < \alpha$ . We present the discrete version as opposed to the continuous analog presented in equation (5). Define the Sinkhorn distance as

$$d_{M,\alpha} := \min_{P \in U_{\alpha}(r,c)} \langle P, M \rangle \tag{8}$$

Then we can introduce an entropy regularization argument stated in a Lagrangian for  $\lambda > 0$ :

$$d_{M}^{\lambda}(r,c) \coloneqq \langle P^{\lambda}, M \rangle,$$
 where 
$$P^{\lambda} = \operatorname{argmin}_{P \in U(r,c)} \langle P, M \rangle - \frac{1}{\lambda} h(P)$$
 (9)

where  $h(P) = -\sum_{i,j=1}^{d} p_{ij} \log(p_{ij})$  is the entropy of P. Then, Sinkhorn's famed algorithm for finding the minimum,

which we know from the general theory will be found on one of the vertices of the polytope, will serve as a proper approximation tool as seen in Algorithm 1. Here, a main result proved by Cuturi is used which states that the solution  $P^{\lambda}$  is unique and, moreover, has the particular form of  $P^{\lambda} = diag(u)Kdiag(v)$ , where u, v are two nonnegative vectors that are unique up to constants and  $K = e^{-\lambda M}$  denotes the matrix exponential of  $-\lambda M$ . This result is pivotal in further speeding up the computation of (9). This type of result is also commonly used as in, for example, [38] which is explored in (6.1.3).

## Algorithm 1: Computation of Entropy-Regularized

```
\begin{aligned} &d = [d_M^\lambda(r,c_1),d_M^\lambda(r,c_2),...,d_M^\lambda(r,c_N)] \\ &\text{Input: } M,\lambda,r,C = [c_1,...,c_N] \\ &I = (r>0); r = r(I); M = M(I,:); K = exp(-\lambda M) \end{aligned}
u = ones(length(r), N)/length(r)
\hat{K} = diag(1./r)K
While u changes or any stopping criterion Do
    u = 1./(\hat{K}(C./(K^Tu)))
end while
v = C./(K^T u)
d = sum(u. * ((K. * M)v))
```

The implementation of Sinkhorn's algorithm to find optimal transport maps has improved the general tractability OT algorithms. We note the improvement on the problem of computing barycenters in Wasserstein space made by Cuturi and Doucet in [11] where they prove the polyhedral convexity of a function that is like a discrete version of (7)

$$f(r, X) = \frac{1}{N} \sum_{i=1}^{N} d(r, c_i, M_{XY_i})$$

where  $d(\cdot, \cdot)$  is the previously defined Sinkhorn distance (8), and r, c are the marginal probabilities.  $M_{XY}$  is the pairwise distance matrix. Here, the problem of the optimal p is phrased using the dual linear programming from known as the dual optimal transport problem:

$$d(r, c, M) = \max_{(\alpha, \beta) \in C_M} \alpha^T r + \beta^T c$$

where  $C_M$  is the polyhedron of dual variables

$$C_M = \{(\alpha, \beta) \in \mathbb{R}^{n+m} | \alpha_i + \beta_j \le m_{ij} \}$$

This problem then has a solution and the computation of the barycenters centers around this.

While the theoretical groundwork for optimal transport has been laid, efficient algorithms are still needed for it to be implemented in large scale. Genevay, et al. formulate stochastic descent methods for large scale computations, making use of the duality arguments previously presented along with entropic regularization for various cases in [15]. Then, Sinkhorn's algorithm will play an important role in the discrete case while the continuous case is very elegantly dealt with using reproducing kernel Hilbert spaces.

For a complete discussion of the numerical methods associated with the OT problem as well as other relevant algorithms, see [40, 23].

## 6 Applications

Here, we hope to bring light to some of the many applications of OT within a machine learning setting.

#### 6.1 Computer Vision

OT finds a natural formulation within the context of computer vision. The common method is to make a probability measure out of color histograms relating to the image. Then, one can find a dissimilarity measure between the images using the Wasserstein distance. An early formulations of OT in computer visions can be in [30] and those relating to the Earth Mover's Distance (EMD) which acts as a slighty different discrete version of the 1-Wasserstein distance. A formulation of the EMD on discrete surfaces can be found in [37]. In the forthcoming, we note a use of OT in improving GANs and the Convolutional Wasserstein Distances which serve well for image interpolation.

#### 6.1.1 OT Meets GANs

Multiple attempts have been made to improve GANs using optimal transport. Arjovski *et al.* recast the GANs problem into an OT theory problem [4]. OT lends itself well to the GANs problem of learning models that generate data like images or text with a distribution that is similar to that of training data. Here in WGANs, we can take two probability measures  $\mu_0, \mu_1 \in M$  with  $\mu_1$  being the distribution of a locally Lipschitz  $g_{\theta}(Z)$  acting as a neural network with *nice* convergence properties and with Z a random variable with density  $\rho$  and  $g_{\theta}$  and the Kantorovich-Rubinstein duality gives

$$W(\mu_0, \mu_1) = \sup_{\|f\| \le 1} \mathbb{E}_{x \sim \mu_0}[f(x)] - \mathbb{E}_{x \sim \mu_1}[f(x)]$$

with supremum taken over all Lipschitz continuous functions  $f:M\to\mathbb{R}$ . It is shown here that there is a solution to this problem with relevant gradient

$$\nabla_{\theta} W(\mu_0, \mu_1) = -\mathbb{E}_{z \sim \rho} [\nabla_{\theta} f(g_{\theta}(z))]$$

wherever both are well-defined. This formulation poses an alternative to the classical GANs and it is found to be more stable, specially when dealing with lower dimensional data, than its counterparts.

We also note the progress made by Salimans *et al.* [31] where they improve upon the idea of mini-batches [14] and using energy functionals [5] to introduce an OT variant using the W-distance named the Minibatch Energy Distance:

$$D_{MED}^{2}(\mu_{0}, \mu_{1}) = 2\mathbb{E}[W_{c}(X, Y)] - \mathbb{E}[W_{c}(X, X')] - \mathbb{E}[W_{c}(Y, Y')]$$

where X, X' are sampled mini-batches from  $\mu_0$  and Y, Y' are sampled mini-batches from  $\mu_1$  and c is the optimal transport function that is learned adversarially through the alternating gradient descent common to GANs. These algorithms are seeing a greater statistical consistency.

#### **6.1.2** Semantic Correspondence

OT is one of the few, if not the only, method that deals with mass-splitting phenomenon which commonly occurs in establishing dense correspondence across semantically similar images. This occurrence is in the form of a many-to-one matching in the assignment of pixels from a source of pixels to a target pixel as well as a one-to-many matching of the same type. The one-to-one matching problem can be recast as an OT problem as done in [21]. Liu *et al.* replace it with maximizing a total correlation where the optimal matching probability is denoted as

$$P^* = \operatorname{argmax}_P \sum_{i,j} P_{ij} C_{ij}$$

where  $P \in \mathbb{R}_+^{n \times m}$ ,  $P\mathbb{1}_n = r$ ,  $P^T\mathbb{1}_m = c$  and r, c are marginals in the same vein as in the section on computational challenges. Then, we can call M = 1 - C to be the cost matrix. Then, the problem becomes the optimal transport problem

$$P^* = \operatorname{argmin}_P \sum_{i,j} PijM_{ij}$$

where  $P \in \mathbb{R}^{n \times m}_+$ ,  $P\mathbb{1}_n = r$ ,  $P^T\mathbb{1}_m = c$ . This problem can then be solved using known algorithms, like those proposed in the computation challenges section. Using the percentage of correct keypoints (PCK) evaluation metric, their proposed algorithm outperformed state-of -the-art algorithms by 7.4 (or 26%), making it a huge improvement over other methods.

#### 6.1.3 Convolutional Wasserstein Distances

In [36], Solomon *et al.* propose an algorithm for approximating optimal transport distances across geometric domains. Here, they make use of the entropy-regularized Wasserstein distance given by (5) for its computational advantages discussed in the Computational Challenges section:

$$W_{2,\gamma}^{2}(\mu_{0},\mu_{1}) = \inf_{\pi \in \Pi(\mu_{0},\mu_{1})} \left[ \int_{M \times M} d(x,y)^{2} d\pi(x,y) - \gamma H(\pi) \right]$$
(10)

They use Varadhan's formula [39] to approximate the distance d(x, y) by transferring heat from x to y over a short time interval:

$$d(x,y)^2 = \lim_{t\to 0} [-2t \ln H_t(x,y)]$$

where  $H_t$  is the heat kernel associated to the geodesic distance d(x,y). Then, we can use this value in a kernel defined by  $K_{\gamma}(x,y) = e^{-\frac{d(x,y)^2}{\gamma}}$ . We can conclude through algebraic manipulations that

$$W_{2,\gamma}^{2}(\mu_{0},\mu_{1}) = \gamma[1 + \min_{\pi \in \Pi} KL(\pi|K_{\gamma})]$$

where KL denotes the K-L divergence (6). Then, in order to compute the convolutional distances, we can discretize the domain M with function and density vectors  $\mathbf{f} \in \mathbb{R}^n$ . Then, define area weights vector  $\mathbf{a} \in \mathbb{R}^n_+$  with  $\mathbf{a}^T \mathbb{1} = 1$  and a symmetric matrix  $\mathbf{H}_t$  discretizing  $H_t$  such that

$$\int_{M} f(x)dx \approx \mathbf{a}^{T} \mathbf{f} \quad \text{and} \quad \int_{M} f(y)H_{t}(\cdot, y)dy \approx \mathbf{H}_{t}(\mathbf{a} \otimes \mathbf{f})$$

Thus we are ready to compute the convolutional Wasserstein distance as in Algorithm 2.

#### Algorithm 2: Convolutional Wasserstein Distance

```
Input: \mu_0, \mu_1, H_t, a, \gamma

Sinkhorn Iterations:

\mathbf{v}, \mathbf{w} \leftarrow 1

for i = 1, 2, 3, ...

\mathbf{v} \leftarrow \mu_0./\mathbf{H}_t(\mathbf{a}. * \mathbf{w})

\mathbf{w} \leftarrow \mu_1./\mathbf{H}_t(\mathbf{a}. * \mathbf{v})

KL Divergence:

Return \gamma \mathbf{a}^t[(\mu_0. * \ln(\mathbf{v})) + (\mu_1. * \ln(\mathbf{w})]
```

We note the authors' use of the Convolution Wasserstein Distance along with barycenters in the Wasserstein space to implement an image interpolation algorithm.

### 6.2 Graphs

The OT problem also lends itself to the formulation of dissimilarity measures within different contexts. In [19], authors developed a fast framework, referred to as WEGL (Wasserstein Embedding for Graph Learning), to embed graphs in a vector space.

We find that analogs of the dissimilarity measures can be defined on graphs and manifolds where the source manifold and target manifolds need not be the same. In [43], Xu et al. propose a new method to solve the joint problem of learning embeddings for associated graph nodes and graph matching. This is done using a regularized Gromov-Wasserstein discrepancy when computing the levels of dissimilarity between graphs. The computed distance allows us to study the topology each of the spaces. The Gromov-Wasserstein discrepancy was proposed by Peyre as a succession to the Gromov-Wasserstein distance which is defined as follows:

**Definition:** Let  $(X, d_X, \mu_X)$  and  $(Y, d_Y, \mu_Y)$  be two metric measure spaces, where  $(X, d_X)$  is a compact metric space and  $\mu_X$  is a probability measure on X (with  $(Y, d_Y, \mu_Y)$  defined in the same way). The Gromov

Wasserstein distance  $d_{GW}(\mu_X, \mu_Y)$  is defined as

$$\inf_{\pi \in \Pi(\mu_X, \mu_Y)} \int_{X \times Y} \int_{X \times Y} L(x, y, x', y') d\pi(x, y) d\pi(x', y'),$$

where  $L(x, y, x', y') = |d_X(x, x') - d_Y(y, y')|$  is the loss function and  $\Pi(\mu_X, \mu_Y)$  is the set of all probability measures on X x Y with  $\mu_X$  and  $\mu_Y$  as marginals. We note that the loss function could be continuous depending on the topology the metric space X is endowed with. At the very least, we would want it to be  $\pi$ -measurable.

When  $d_x$  and  $d_y$  are replaced with dissimilarity measurements rather than strict distance metrics and the loss function L is defined more flexibly, the GW distance can be relaxed to the *discrepancy*. From graph theory, a graph is represented by its vertices and edges, G(V,E). If we let a metric-measure space be defined by the pair  $(C,\mu)$ , then we can define the Gromov-Wasserstein discrepancy between two spaces,  $(C_s, \mu_s)$  and  $(C_t, \mu_t)$ , as:

$$d_{GW}(\mu_s, \mu_t) = \min_{T \in \pi(\mu_s, \mu_t)} \sum_{i,j,i',j'} L(c_{ij}^s, c_{i'j'}^t) Tii' Tjj'$$
$$= \min_{T \in \pi(\mu_s, \mu_t)} \langle L(C_s, C_t, T), T \rangle$$

In order to learn the mapping that includes the correspondence between graphs and also the node embeddings, Xu et al. proposed the regularized GW discrepancy:

$$\min_{X_s, X_t} \min_{T \in \Pi(\mu_s, \mu_t)} \langle L(C_s(X_s), C_t(X_t), T), T \rangle + \alpha \langle K(X_s, X_t), T \rangle + \beta R(X_s, X_t)$$

To solve this problem, the authors present Algorithm 3.

## Algorithm 3: Gromov-Wasserstein Learning (GWL)

```
Input: \{C_s, C_t\}, \{\mu_s, \mu_t\}, \beta, \gamma, the dimension D, the number of outer/inner iterations \{M, N\}. Output: X_s, X_t, and \hat{T}
Initialize X_s^{(0)}, X_t^{(0)} randomly, \hat{T}^{(0)} = \mu_s \mu_t^T. For m = 0: M - 1:
Set \alpha_m = \frac{m}{M}.
For n = 0: N - 1
Update optimal transport \hat{T}^{(m+1)}
Obtain X_s^{(m+1)}, X_t^{(m+1)}
X_s = X_s^{(M)}, X_t = X_t^{(M)} and \hat{T} = \hat{T}^{(M)}. Graph matching: Initialize correspondence set P = \emptyset
For v_i \in V_s
j = \operatorname{argmax}_j \hat{T}_{ij}, P = P \bigcup \{(v_i \in V_s, v_j \in V_t)\}.
```

The proposed methodology produced matching results that are better than all other comparable methods and opens the opportunity for the improvement of well-known systems (i.e. recommendation systems).

We note that the Gromov-Wasserstein discrepancy can also be used to improve GANs, as is done in [7]. Here, Bunne, et al., adapt the generative model to use the Gromov-Wasserstein discrepancy to perform GANs across different types of data.

#### 6.3 Neural Architecture Search

In this section we will look at the following paper: Neural Architecture Search with Bayesian Optimisation and Optimal Transport [18]. Bayesian Optimization (BO) refers to a set of methods used for optimization of a function f, thus making it perfect for solving the model selection problem over the space of neural architectures. The difficulty posed in BO when dealing with network architecture is figuring out how to quantify (dis)similarity between any two networks. To do this, the authors developed what they call a (pseudo-)distance for neural network architectures, called OTMANN (Optimal Transport Metrics for Architectures of Neural Networks). Then, to perform BO over neural

network architectures, they created NASBOT, or Neural Architecture Search with Bayesian Optimization and Optimal Transport. To understand their formulation, we first look at the following definitions and terms. First, a Gaussian process is a random process characterized by an expectation function (mean function)  $\mu:\chi\to\mathbb{R}$  and a covariance (kernel)  $\kappa=\chi^2\to\mathbb{R}$ . In the context of architecture search, having a large  $\kappa(x,x')$ , where  $x,x'\in\chi$  and  $\kappa(x,x')$  is the measure of similarity so that f(x) and f(x') are highly correlated; implying the GP imposes a smoothness condition on  $f:\chi\to\mathbb{R}$ . Next, the authors view a neural network (NN) as a graph whose vertices are the layers of the network G=(L,E), where L is a set of layers and E the directed edges. Edges are denoted by a pair of layers,  $(u,v)\in E$ . A layer  $u\in L$  is equipped with a layer label ll(u), which denotes the type of operations performed at layer u (i.e. ll(1)=conv3 means 3x3 convolutions). Then, the attribute lu denotes the number of computational units in a layer. Furthermore, each network has  $decision\ layers$ , which are used to obtain the predictions of the network. When networks have more than one decision layer, one considers the average of the output given by each layer. Lastly, each network has an input and output layer,  $u_{in}$  and  $u_{op}$  respectively; any other layer is denoted as a  $decision\ layer$ .

Using the definitions above, the authors describe the distance for neural architectures as  $d:\chi^2\to\mathbb{R}_+$ ; with the goal of obtaining a kernel for the GP where  $\kappa(x,x')=exp(-\beta d(x,x')^p)$ , given that  $\beta,p\in\mathbb{R}_+$ . We first look at the OTMANN distance. OTMANN is defined as the minimum of a matching scheme which attempts to match the computation at the layers of one network to the layers of another, where penalties occur given that different types of operations appear in matched layers. The OTMANN distance is that which minimizes said penalties. Given two networks  $G_1(L_1,E_1)$  and  $G_2(L_2,E_2)$  with  $n_1,n_2$  layers respectively, the OTMANN distance is computed by solving the following optimization problem:

In the above equation,  $\phi_{lmm}$  is the label mismatch penalty,  $\phi_{str}$  is the structural term penalty,  $\phi_{nas}$  is the non-assignment penalty,  $Z \in \mathbb{R}^{n_1 x n_2}$  denoting hte maount of mass matched between layer  $i \in G_1$  and  $j \in G_2$ ,  $l_m : L \to \mathbb{R}_+$  is a layer mass, and lastly  $\nu_{str} > 0$  determines the trade-off between the structural term and other terms. This problem can be formulated as an Optimal Transport problem and is proved in the appendix of the paper.

Next, we look at NASBOT. The goal here is to use the kernel  $\kappa$ , as previously mentioned, to define the neural architectures and to find a method to optimize the acquisition function:

$$\phi_t(x) = \mathbb{E}[\max\{0, f(x) - \tau_{t-1}\} | \{(x_i, y_i)\}_{i=1}^{t-1}],$$
  
$$\tau_{t-1} = \underset{i \le t-1}{\operatorname{argmax}} f(x_i)$$

The authors solve this optimization problem using an evolutionary algorithm, whose solution leads to the creation of NASBOT. Detailed explanations on the algorithm and the methodology onto which the optimization was solved can be found in the appendix of the original paper. After running an experiment to compare NASBOT against known methods, the authors show that NASBOT consistently had the smallest cross validation mean squared error. For the interested reader, there are illustrations for the best architectures found for the problem posed in the experiment proposed.

## 6.4 Document Representation

In this section we will look at the following paper: Hierarchical Optimal Transport for Document Representation [45]. In this paper, Yurochkin, et al. combine hierarchical latent structures from topic models with geometry from word embeddings. Hierarchical optimal topic transport document distances, referred to as HOTT, this method combines language information (via word embeddings) with topic distributions from latent Dirichlet allocation (LDA) to measure the similarities between documents. Given documents  $d^1$  and  $d^2$ , HOTT is defined as:

$$HOTT(d^{1}, d^{2}) = W_{1}(\sum_{k=1}^{|T|} \bar{d}_{k}^{1} \delta_{t_{k}}, \sum_{k=1}^{|T|} \bar{d}_{k}^{2} \delta_{t_{k}})$$
(11)

Here,  $\bar{d}^i$  represents document distributions over topics and the Dirac delta  $\delta_{t_k}$  is a probability distribution supported on the corresponding topic  $t_k$  and  $W_1(d^1,d^2) = WMD(d^1,d^2)$  (WMD being the Word Movers Distance). By truncating topics, the authors were able to reduce the computational time and make HOTT a competitive model against common methods. Their experiments show that although there is no uniformly best method, HOTT has on average the smallest error with respect to nBOW (normalized bag of words). More importantly, what was shown was that the process of truncating topics to improve computational time does not hinder the goal of obtaining high-quality distances. Interested readers will find in the paper more detailed reports about the setup and results of the experiments run.

#### 6.5 Domain Adaptation

In this section we will cover *Optimal Transport for Domain Adaptation* [12]. In their paper, Flamary, *et al.*, propose a regularized unsupervised optimal transportation model to perform an alignment of the representations in the source and target domains. By learning a transportation plan that matches the source and target PDFs, they constrained labeled samples of the same class during the transport. This helps solve the discrepancies (known as drift) in data distributions.

In real world problems, the drift that occurs between the source and target domains generally implies a change in marginal and conditional distributions. In this paper, the authors assume the domain drift is due to "an unknown, possibly nonlinear transformation of the input space  $T:\Omega_s\to\Omega_t$  (omega is a measurable space, s is source, t is target). Because searching for T is an intractable problem and requires restrictions to become approximated. Here, the authors consider the problem of finding T the same as choosing a T such that one minimizes the transportation cost C(T):

$$C(T) = \int_{\Omega_s} c(x, T(x)) d\mu(x)$$
(12)

where  $c: \Omega_s x \Omega_t \to \mathbb{R}^+$  and  $\mu(x)$  is a probability mass (or measure from x to T(x).)

This is precisely the optimal transport problem. Then, to further improve the computational aspect of the model, a regularization component that preserves label information and sample neighborhood during the transportation is introduced. Now, the problem is as follows:

$$\min_{\pi \in \Pi} \langle \pi, C \rangle_F + \lambda \Omega_s(\pi) + \eta \Omega_c(\pi)$$
(13)

where  $\lambda \in \mathbb{R}$ ,  $\eta \geq 0$ ,  $\Omega_c(\cdot)$  is a class-based regularization term, and

$$\Omega_s(\pi) = \sum_{i,j} \pi(i,j) \log(i,j)$$

This problem is solved using Algorithm 4:

## **Algorithm 4: Generalized Conditional Gradient**

Initialize: k = 0, and  $\pi^0 \in P$  repeat

With  $C \in \nabla f(\pi^k)$  solve  $\pi$ 

With  $G \in \nabla f(\pi^k)$ , solve  $\pi^* = \operatorname*{argmin}_{\pi \in B} \langle \pi, G \rangle_F + g(\pi)$ Find the optimal step  $\alpha^k$ ,  $\alpha^k = \operatorname*{argmin}_{0 \leq \alpha \leq 1} f(\pi^k + \alpha \Delta \pi) + g(\pi^k + \alpha \Delta \pi)$ , with  $\Delta \pi = \pi^* - \pi^k$ 

$$\pi^{k+1} \leftarrow \pi^k + \alpha^k \Delta \pi$$
, set  $k \leftarrow k+1$ 

until Convergence

In the algorithm above,  $f(\pi) = \langle \pi, C \rangle_F + \eta \Omega_c(\pi)$  and  $g(\pi) = \lambda \Omega_s(\pi)$ . Using the assumption that  $\Omega_c$  is differentiable, step 3 of the algorithm becomes

$$\pi^* = \underset{\pi \in Pi}{\operatorname{argmin}} \langle \pi, C + \eta \nabla \Omega_c(\pi^k) \rangle_F + \lambda \Omega_s(\pi)$$

By using a constrained optimal transport method, the overall performance was better than other state-of-the-art methods. Readers can find detailed reports on Table 1 in [12].

For readers interested in domain adaptation, a varying approach to study heterogeneous domain adaptation problems using OT can be found in [44].

#### 7 Future Research

Further research will allow OT to be implemented in more areas and become more widely acceptable. The main problem with optimal transport is scaling onto higher dimensions. The optimal mappings that need to be solved are currently intractable in high-dimensions, which is where most of the current problems today lie. For example, Google's NLP model has roughly 1 trillion parameters. This type of problem is currently outside the scope of OT.

Another interesting research topic is the use of optimal transport in approximating intractable distributions. This would compete with current known methods like KL-divergence and open up interesting opportunities when working with variational inference and/or expectation propagation. Another fundamental area to explore lies with the choice of using the Wasserstein distance. As shown throughout the paper, it is the most commonly used metric, but as one can see in Appendix 1, there are various others metrics, or distances, that may be used to replace W-distance. Interested readers can read more about them in Villani's Book, *Optimal transport: old and new* [41]. For further research from an applied perspective, one possibility is the use of the GWL framework explained in section 6.2 to improve on recommendation systems. On the other hand, all of the papers we have referenced above are quite novel in their applications and thus they all provide space for continuation or extension into more specific sub-fields within their respective context.

## **8 Concluding Remarks**

Throughout this survey, we have shown that Optimal Transport is seeing growing attention within the machine learning community due to its applicability in different areas. Although OT is becoming widely accepted in the machine learning world, it is deeply rooted in mathematics and so we extracted the most important topics so that interested readers can access only what is needed to have a high-level understanding of what is happening. These excerpts explain Kantorovich duality, entropic regularization, KL divergence, and Wasserstein barycenters. Although the applications of OT span a wide range, it is limited by computational challenges. Within this section we explored how using an entropic regularization term allowed for the formation of an algorithm that made OT problems computationally feasible and thus applicable. This takes us to the last section of this survey, the applications of optimal transport in machine learning. We began with computer vision, as it was one of the first applications of OT in ML. First, OT has been used to improve GANs by providing better statistical stability in low-dimensional data. Furthermore, since OT is one of the few methods that deal with the mass-splitting phenomenon, it allowed for many-to-one matching in pixel assignments which yielded a new approach to semantic correspondence with a 26% performance improvement over state-of-the-art methods. The last application we covered with respect to computer vision was the use of W-distance to create a novel method for image interpolation called Convolutional Wasserstein Distance. Next, with respect to graphs, OT has allowed for the creation of the Gromov-Wasserstein Learning (GWL) algorithm which have also been shown to improve GANs. Other interesting areas that OT has shown promising results include neural architecture search, document representation, and domain adaptation. All of the papers we have analyzed and summarized will show that in some form (computational/accuracy) the use of OT has yielded better results than traditional methods. Although the computational inefficiencies are prevalent, the future for optimal transport in machine learning looks promising as more researchers become aware of this new intersection of areas.

## Appendix A

The implementation of the conclusions of OT in machine learning rely mostly on the implementation of the various metrics that can be used as error measures in model tuning. The most notable ones arise from the reformulation or approximation of metrics into convex functionals that can be optimized by drawing on the many beautiful conclusions of convex geometry. Here, we recall a metric, many times called a distance, as a function  $d(\cdot, \cdot): X \times X \to \mathbf{R}_+$ , where X is a metrizable space, that satisfies

- Positive Definite:  $d(x,y) \ge 0$  with d(x,y) = 0 if, and only if, x = y
- Symmetric: d(x, y) = d(y, x) for all  $x, y \in X$
- Subadditive:  $d(x,y) \le d(x,z) + d(z,y)$  for all  $x,y,z \in X$

Here, we want to note some different error estimates that come up in the OT literature as well as some that are traditionally used to compare probability distributions. The most notable comparison of probability measures in the OT literature is the p-Wasserstein Distance

$$W_p^d(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0, \mu_1)} \left( \int_{M \times M} d(x, y)^p d\pi(x, y) \right)^{1/p}$$
(14)

In 14, d is a metric. We see from definition that it should very much act like a minimal  $L^p$  distance on the space of probability measures. The most relevant choice of parameter p is p=1,2. This distance was formulated in the most general sense possible and it has a natural discrete formulation for discrete measures. Therefore, it allows for different contexts. For example, we saw the analog in the context of graphs as the Gromov-Wasserstein distance as:

Let  $(X, d_X, \mu_X)$  and  $(Y, d_Y, \mu_Y)$  be two metric measure spaces, where  $(X, d_X)$  is a compact metric space and  $\mu_X$  is a probability measure on X (with  $(Y, d_Y, \mu_Y)$  defined in the same way). The Gromov-Wasserstein distance  $d_{GW}(\mu_X, \mu_Y)$  is defined as

$$\inf_{\pi \in \Pi(\mu_X, \mu_Y)} \int_{X \times Y} \int_{X \times Y} L(x, y, x', y') d\pi(x, y) d\pi(x', y'), \tag{15}$$

where  $L(x, y, x', y') = |d_X(x, x') - d_Y(y, y')|$ . Here, we see that the formulas look naturally similar. The Gromov-Wasserstein distance would be a particular choice of the 1-Wasserstein distance to a general metric space which can then be relaxed to be able to work with graphs as we saw before.

The novelty in using OT in applications is principally the different error estimates. We recall some of the well-known distances that are traditionally used to compare probability measures:

• KL Divergence:

$$\begin{split} KL(\pi|\kappa) &\coloneqq \\ &\int \int_{M\times M} \pi(x,y) \bigg[ \ln \frac{\pi(x,y)}{\kappa(x,y)} - 1 \bigg] dx dy \end{split}$$

- Hellinger distance:  $H^2(\mu_0, \mu_1) = \frac{1}{2} \int (\sqrt{\frac{d\mu_0}{d\lambda}} \sqrt{\frac{d\mu_1}{d\lambda}})^2 d\lambda$ , where  $\mu_0, \mu_1$  are absolutely continuous with respect to  $\lambda$  and  $\frac{d\mu_0}{d\lambda}, \frac{d\mu_1}{d\lambda}$  denote the Radon-Nykodym derivatives, respectively.
- Lèvy-Prokhorov distance:  $d_P(\mu_0, \mu_1) = \inf\{\epsilon > 0; \exists X, Y; \inf \mathbb{P}[d(X, Y) > \epsilon] \leq \epsilon\}$
- Bounded Lipschitz distance (or Fortet-Mourier distance):  $d_b L(\mu_0, \mu_1) = \sup\{\int \phi d\mu_0 \int \phi d\mu_1; ||\phi||_{\infty} + ||\phi||_{\text{Lip}} \leq 1\}$
- (in the case of nodes) Euclidean distance:  $d(x,y) = \sqrt{(x-y)^2}$

We note that the Lèvy-Prokhorov and bounded Lipschitz distances can work in much the same way that the Wasserstein distance does. At the present, the Wasserstein distance proves useful because of it's capabilities in dealing with large distances and its convenient formulation in many problems such as the ones presented in this paper as well as others coming from partial differential equations. It's definition using infimum makes it easy to majorate. Its duality properties are useful–particularly in the case when p=1 as we see with the Kantorovich-Rubinstein distance where it is defined as an equivalence to its dual:

$$W_1(\mu_0, \mu_1) = \sup_{\|\phi\|_{\text{Lip}} \le 1} \left\{ \int_X \phi d\mu_0 - \int_x \phi d\mu_1 \right\}$$
 (16)

The interested reader can read more about the different distances in [27, 41]

As we presently see in this paper, we notice that much of the work on the optimal transport in machine learning is in the reformulation of the algorithms, which classically used the traditional distances, into new versions that use the Wasserstein distance. Then, a lot of the work is done in dealing with the computational inefficiency of the Wasserstein distance. Moving forward, the authors think that many machine learning algorithms will implement some of the "deeper" features of the optimal transport theory to improve such algorithms after their best formulation becomes abundantly clear.

## Appendix B

For the readers interested in papers that apply OT in machine learning, here are a few more references to be considered. First we have OT in GANS:

• A geometric view of optimal transportation and generative model [20]

Next, for semantic correspondence and NLP we have:

• Improving sequence-to-sequence learning via optimal transport [8]

Lastly, on domain adaptation we have:

- *Joint distribution optimal transportation for domain adaptation* [9]
- Theoretical analysis of domain adaptation with optimal transport [28]

## Acknowledgement

This material is based upon Luiz Manella Pereira's work supported by the U.S. Department of Homeland Security under Grant Award Number, 2017-ST-062-000002. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

#### References

- [1] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.
- [2] Najma Ahmad. "The geometry of shape recognition via the Monge-Kantorovich optimal transport problem." In: (2003).
- [3] Luca Ambrogioni et al. "Wasserstein variational inference". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2473–2482.
- [4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein gan". In: *arXiv preprint arXiv:1701.07875* (2017).
- [5] Marc G. Bellemare et al. "The Cramer Distance as a Solution to Biased Wasserstein Gradients". In: *CoRR* abs/1705.10743 (2017). arXiv: 1705.10743. URL: http://arxiv.org/abs/1705.10743.
- [6] Patrick Bernard and Boris Buffoni. "Optimal mass transportation and Mather theory". In: *arXiv preprint math/0412299* (2004).
- [7] Charlotte Bunne et al. "Learning generative models across incomparable spaces". In: *arXiv preprint* arXiv:1905.05461 (2019).
- [8] Liqun Chen et al. "Improving sequence-to-sequence learning via optimal transport". In: *arXiv preprint* arXiv:1901.06283 (2019).
- [9] Nicolas Courty et al. "Joint distribution optimal transportation for domain adaptation". In: *arXiv preprint* arXiv:1705.08848 (2017).
- [10] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems*. 2013, pp. 2292–2300.
- [11] Marco Cuturi and Arnaud Doucet. "Fast computation of Wasserstein barycenters". In: (2014).
- [12] R Flamary et al. "Optimal transport for domain adaptation". In: IEEE Trans. Pattern Anal. Mach. Intell (2016).

- [13] Wilfrid Gangbo and Robert J McCann. "Shape recognition via Wasserstein distance". In: *Quarterly of Applied Mathematics* (2000), pp. 705–737.
- [14] Aude Genevay, Gabriel Peyré, and Marco Cuturi. *Learning Generative Models with Sinkhorn Divergences*. 2017. arXiv: 1706.00292 [stat.ML].
- [15] Aude Genevay et al. "Stochastic optimization for large-scale optimal transport". In: *Advances in neural information processing systems*. 2016, pp. 3440–3448.
- [16] Steven Haker and Allen Tannenbaum. "On the Monge-Kantorovich problem and image warping". In: *IMA Volumes in Mathematics and its Applications* 133 (2003), pp. 65–86.
- [17] Steven Haker et al. "Optimal mass transport for registration and warping". In: *International Journal of computer vision* 60.3 (2004), pp. 225–240.
- [18] Kirthevasan Kandasamy et al. "Neural architecture search with bayesian optimisation and optimal transport". In: *Advances in neural information processing systems*. 2018, pp. 2016–2025.
- [19] Soheil Kolouri et al. "Wasserstein Embedding for Graph Learning". In: arXiv preprint arXiv:2006.09430 (2020).
- [20] Na Lei et al. "A geometric view of optimal transportation and generative model". In: *Computer Aided Geometric Design* 68 (2019), pp. 1–21.
- [21] Yanbin Liu et al. "Semantic Correspondence as an Optimal Transport Problem". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 4463–4472.
- [22] John N Mather. "Minimal measures". In: Commentarii Mathematici Helvetici 64.1 (1989), pp. 375–394.
- [23] Quentin Merigot and Boris Thibert. "Optimal transport: discretization and algorithms". In: *arXiv preprint* arXiv:2003.00855 (2020).
- [24] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences de Paris* (1781).
- [25] Shmuel Peleg, Michael Werman, and Hillel Rom. "A unified approach to the change of resolution: Space and gray-level". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.7 (1989), pp. 739–742.
- [26] Gabriel Peyré, Marco Cuturi, et al. "Computational optimal transport". In: *Foundations and Trends*® *in Machine Learning* 11.5-6 (2019), pp. 355–607.
- [27] S.T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley Series in Probability and Statistics Applied Probability and Statistics Section. Wiley, 1991. ISBN: 9780471928775. URL: https://books.google.com/books?id=5grvAAAAMAAJ.
- [28] Ievgen Redko, Amaury Habrard, and Marc Sebban. "Theoretical analysis of domain adaptation with optimal transport". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2017, pp. 737–753.
- [29] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. "A metric for distributions with applications to image databases". In: *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE. 1998, pp. 59–66.
- [30] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. "The earth mover's distance as a metric for image retrieval". In: *International journal of computer vision* 40.2 (2000), pp. 99–121.
- [31] Tim Salimans et al. "Improving GANs using optimal transport". In: arXiv preprint arXiv:1803.05573 (2018).
- [32] Roman Sandler and Michael Lindenbaum. "Nonnegative matrix factorization with earth mover's distance metric for image analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8 (2011), pp. 1590–1602.
- [33] Filippo Santambrogio. "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55.58-63 (2015), p. 94.
- [34] Filippo Santambrogio. "Optimal Transport meets Probability, Statistics and Machine Learning". In: ().
- [35] Meyer Scetbon et al. "Handling multiple costs in optimal transport: Strong duality and efficient computation". In: *arXiv preprint arXiv:2006.07260* (2020).
- [36] Justin Solomon et al. "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains". In: *ACM Transactions on Graphics (TOG)* 34.4 (2015), pp. 1–11.
- [37] Justin Solomon et al. "Earth mover's distances on discrete surfaces". In: *ACM Transactions on Graphics (TOG)* 33.4 (2014), pp. 1–12.
- [38] Justin Solomon et al. "Wasserstein propagation for semi-supervised learning". In: *International Conference on Machine Learning*. 2014, pp. 306–314.
- [39] Sathamangalam R Srinivasa Varadhan. "On the behavior of the fundamental solution of the heat equation with variable coefficients". In: *Communications on Pure and Applied Mathematics* 20.2 (1967), pp. 431–455.

- [40] François-Xavier Vialard. "An elementary introduction to entropic regularization and proximal methods for numerical optimal transport". In: (2019).
- [41] Cédric Villani. Optimal transport: old and new. Vol. 338. Springer Science & Business Media, 2008.
- [42] Cédric Villani. Topics in optimal transportation. 58. American Mathematical Soc., 2003.
- [43] Hongteng Xu et al. "Gromov-wasserstein learning for graph matching and node embedding". In: *arXiv preprint arXiv:1901.06003* (2019).
- Yuguang Yan et al. "Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation." In: *IJCAI*. 2018, pp. 2969–2975.
- [45] Mikhail Yurochkin et al. "Hierarchical optimal transport for document representation". In: *Advances in Neural Information Processing Systems*. 2019, pp. 1601–1611.