Optimal Transport in Statistical Machine Learning: Selected Review and Some Open Questions

Abhinav Maurya
Carnegie Mellon University
Pittsburgh, PA 15213
ahmaurya@cmu.edu

Abstract

We review recent theoretical results underpinning the use of optimal transport and Wasserstein distances in statistical machine learning. The four primary results described in this report relate to the use of Wasserstein loss for train multi-class or multi-label classifiers [1], the use of composite Wasserstein loss to establish the convergence and contraction rates for mixing distributions in certain finite and infinite mixture models [2], the guarantees provided by optimal transport when used for domain adaptation [3, 4], and the use of Wasserstein loss in improving the training procedure of Generative Adversarial Networks [5].

1 Introduction

Optimal mass transport is an elegant mathematical tool at the intersection of probability theory and mathematical optimization. Formalized by Monge in 1781 [6], a generalized version of Monge's problem was stated by Kantorovich in 1942 [7]. However, theoretical results about the existence of Monge's optimal transport maps and Kantorovich's optimal transport plans (and the accompanying regularity conditions) have been worked out relatively recently [8, 9]. Spurred by research on the efficient computation of optimal transport maps [10] and plans [11] and the association between optimal transport and the Wasserstein metric, optimal transport has been applied in formulating solutions to numerous machine learning problems such as learning document distances [12, 13], MCMC-free sampling from Bayesian posteriors [14], image retrieval [15], histogram regression [16], domain adaptation [17, 3, 4], kernel/metric learning [18, 19], multi-label classification [1], label distribution learning [20], improved training of Generative Adversarial Networks (GANs) [5, 21], etc. with impressive results. The associated Wasserstein metric has also been used in statistical analyses such as high-dimensional two-sample testing [22] and deriving the convergence and posterior contraction rates for finite and infinite mixture models [2].

In this report, we survey theoretical results that support the methodological constructs based on optimal transport being developed by the machine learning community. We focus primarily on the problem of optimal transport of discrete measures (primarily histograms, softmax output probabilities, and atoms of mixture models). While it may seem restrictive to limit this project to the study of discrete measures, the use of optimal transport in machine learning theory and methods is almost exclusively limited to such measures [1, 12, 13, 16, 23, 24, 2].

2 Notation, Assumptions, and Key Facts about Optimal Transport

Here, we introduce the problem of optimal transport and its use in defining the Wasserstein metric over the space of probability measures. We also specify the notations and assumptions prevalent in the optimal transport and Wasserstein metric literature. Notations specific to certain papers will be introduced in their respective sections later on.

Preprint. Work in progress.

The problem of optimal transportation was formulated by Monge as the discovery of a measurable map T^* that minimizes the cost of transforming ("pushing") probability measure μ to probability measure ν with respect to a pre-specified ground metric $c(x, y) = |y - x|$

$$\begin{aligned} \underset{T}{\text{minimize}} \quad & \int_{\mathbb{R}^d} c(x, T(x)) f(x) dx \\ \text{subject to} \quad & \int_A g(x) dx = \int_{T^{-1}(A)} f(x) dx \end{aligned} \tag{1}$$

The primary limitation of this formulation is that it doesn't allow any dirac-delta probability masses of the first measure to split while being pushed into the other measure. As a result, optimal transport maps do not exist from a discrete measure to a continuous measure, and exist from a discrete measure to another discrete measure under very limited circumstances. This limitation was rectified by Kantorovich who formulated the optimal transport problem as the discovery of a transport plan γ^* over the domain $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d, \mathbb{R}^d) : T_{X\#}\gamma = \mu, T_{Y\#}\gamma = \nu\}$ where $T_X(x, y) = x$ and $T_Y(x, y) = y$ i.e. T_X and T_Y are marginal projection operators.

$$\begin{aligned} \underset{\gamma \in \Pi(\mu,\nu)}{\text{minimize}} \quad & \int_{\mathbb{R}^d, \mathbb{R}^d} c(x, y) d\gamma(x, y) \\ \text{subject to} \quad & T_{X\#}\gamma = \mu, T_{Y\#}\gamma = \nu \end{aligned} \tag{2}$$

In the discrete case, the optimal transport problem can be stated as follows:

$$\begin{aligned} \underset{\gamma \in \Pi(\mu,\nu)}{\text{minimize}} \quad & \int_{\mathcal{K}, \mathcal{K}} c(\kappa_1, \kappa_2) d\gamma(\kappa_1, \kappa_2) \\ \text{subject to} \quad & T_{X\#}\gamma = \mu, T_{Y\#}\gamma = \nu \end{aligned} \tag{3}$$

Since the problem has been stated as the discovery of a joint probability whose marginals equal the two input measures and which minimizes the cost of transforming the first measure into the second one, it is always well-defined even when we discuss the transformation of discrete measures into continuous or discrete measures. The optimal value of the above optimization problems on transportation plans $\gamma \in \Pi(\mu, \nu)$ is known as the p-Wasserstein distance $W_p(\mu, \nu)$ if the cost $c(\cdot, \cdot)$ is given by the p^{th} power $d^p(\cdot, \cdot)$ of metric $d(\cdot, \cdot)$.

2.1 Properties of Wasserstein Spaces

We have stated the optimal transport problem over the domain $(\mathbb{R}^d, \mathbb{R}^d)$ in the continuous case and (κ, κ) in the discrete case. However, the problem is defined over any metric space (Ω, d). We denote the space of probability measures over Ω endowed with the p-Wasserstein metric as $\mathcal{W}_p(\Omega)$. For any $1 \leq p < \infty$, $\mathcal{W}_p(\Omega)$ is compact only iff the underlying ground metric space Ω is compact. If Ω is not bounded, then $\mathcal{W}_p(\Omega)$ is not even locally compact. $\mathcal{W}_\infty(\Omega)$ is neither compact nor locally compact irrespective of whether Ω is bounded or compact. Also, the various p-Wasserstein distances are ordered through an application of Jensen's inequality i.e. $W_p(\mu, \nu) \leq W_q(\mu, \nu)$ if $p \leq q$. If Ω is bounded i.e. $\text{diam}(\Omega) = D$, then $W_q(\mu, \nu) \leq D^{1-p/q} W_p^{p/q}(\mu, \nu)$ for $p \leq q$.

2.2 Geodesics in Wasserstein Spaces

Below we state McCann's linear interpolation theorem [25] which allows for the construction of constant-speed geodesics using optimal transport plans.

Theorem 1. *$\mathcal{W}_p(\Omega)$ is a length space if Ω is a convex domain in $\mathbb{R}^d$. Also, for $\mu, \nu \in \Omega$ and γ^* being the optimal transport plan for the the ground transport cost $c(x, y) = ||x - y||^p$, the curve given by $\mu^{\gamma^*}(s) = (p_s)_{\#}\gamma^*$ where $p_s(\mu, \nu) = x + s(y - x)$ is a constant-speed geodesic from μ to ν. If $p > 1$, all constant-speed geodesics can be expressed in this form. If μ is absolutely continuous, there is only one such geodesic which has the form $\mu(s) = [(1 - s)id + sT]_{\#}\mu$.*

In $\mu(s)$, id is the identity transport map and T is the optimal transport map from μ to ν. This, for $s = 0$, we obtain $\mu(s) = \mu$ and for $s = 1$, we get $\mu(s) = T_{\#}\mu = \nu$. Varying s from 0 to 1

provides a continuous deformative interpolation between μ and ν linear in the Wasserstein space. This property has widely been used in computer graphics and vision for morphing between images, 3D shapes, and point clouds; for interpolating between colormaps or texture profiles; for realistic style transformations; and morphometry-based drug screening, cancer detection, and analyzing galaxy morphologies. See [8] for a more detailed review of applications.

3 Empirical Risk Minimization with Wasserstein Loss

In multi-class or multi-label classification, the output is often a probability distribution over the output space $\mathcal{K}$: $h(x) \in \mathcal{K}$ or $h(x) \in 2^{|\mathcal{K}|}$. It is common to obtain such probabilities using a hypothesis predictor $h_0(x)$ from a base hypothesis space H^0 and then applying the softmax transformation $s(\cdot)$ to it. While such problems are often solved using information-theoretic divergences such as KL divergence when no additional relationship between the labels is available, it is useful to incorporate the *ground metric* of semantic similarity between labels when it is available. This can be done using the exact Wasserstein loss [1]. The resulting empirical risk minimization problem is as follows:

$$h_{\hat{\theta}} = \operatorname*{argmin}_{h_\theta \in \mathcal{H}} \left\{ \hat{\mathbb{E}}_S[W_p^p(h_\theta(x), y)] = \frac{1}{N} \sum_{i=1}^{N} W_p^p(h_\theta(x_i), y_i) \right\} \tag{4}$$

[1] further makes the assumption that $\mathcal{H} = s \circ \mathcal{H}^0$ i.e. the result is obtained by applying the softmax operation s to a base hypothesis $\mathcal{H}^0$ that maps the input into $\mathbb{R}^k$. The Wasserstein distance $W_p^p(h(x), y)$ is the result of an optimal transport problem:

$$\begin{aligned} W_p^p(h(x), y) \quad &= \underset{T \in \Pi(h(x), y)}{\text{minimize}} \quad \langle T, M \rangle \\ &\qquad \text{subject to} \qquad T \in \mathbb{R}_+^{K \times K}, T\mathbf{1} = h(x), T'\mathbf{1} = y \end{aligned} \tag{5}$$

Here, $M \in \mathbb{R}_+^{K \times K}$ is the cost matrix which is equivalent to $M_{\kappa,\kappa'} = d_{\mathcal{K}}^p(\kappa, \kappa')$ calculated from the ground metric $d_{\mathcal{K}}(\cdot, \cdot)$

Since the calculation of Wasserstein distances is a linear programming problem, it is expensive to compute considering that the distance $W_p^p(h(x), y)$ might need to be computed between all pairs of data points in the dataset. Regularizing the above objective using entropy of the plan $H(T)$ makes the objective strictly convex.

$$\begin{aligned} W_{p,\lambda}^p(h(x), y) \quad &= \underset{T \in \Pi(h(x), y)}{\text{minimize}} \quad \langle T, M \rangle - \frac{1}{\lambda} H(T) \\ &\qquad \text{subject to} \qquad T \in \mathbb{R}_+^{K \times K}, T\mathbf{1} = h(x), T'\mathbf{1} = y \end{aligned} \tag{6}$$

The form of the entropy-regularized objective allows a solution through *iterated diagonal scaling* commonly known as the Sinkhorn-Knopp algorithm [11] which involves matrix-matrix or matrix-vector multiplications and therefore can be speeded up easily using parallelized linear algebra libraries or GPGPU programming.

Theorem 2. *For $p = 1$ and any $\delta > 0$, with probability at least $1 - \delta$, the following holds*

$$\mathbb{E}[W_1(h_{\hat{\theta}}(x), y)] \quad \leq \inf_{h_\theta \in \mathcal{H}} \mathbb{E}[W_1(h_\theta(x), y)] + 32KC_M\mathcal{R}_N(\mathcal{H}^0) + 2C_M\sqrt{\frac{\log(1/\delta)}{2N}} \tag{7}$$

Here, $K = |\mathcal{K}|$ is the number of output labels, $C_M = \max(M)$ i.e. the maximum of the distance matrix M and $\mathcal{R}_N(\mathcal{H}^0)$ is the Rademacher complexity of the space $\mathcal{H}^0$ from which the base hypothesis h_0 is learned before a softmax transformation is applied to get output probabilities. For most classifiers such as kernel machines or neural networks, $\mathcal{R}_N(\mathcal{H}^0)$ decreases as the number of datapoints N increases.

The above theorem for multi-label classification leads to similar statistical learning bound for the multi-class setting where only one label is predicted per datapoint.

Theorem 3. *For $\kappa_{\hat{\theta}}(x) = \text{argmax}_{\kappa} h_{\hat{\theta}}(\kappa|x)$, $p = 1$, and any $\delta > 0$, with probability at least $1 - \delta$, the following holds*

$$\mathbb{E}_{x,\kappa}[d_{\kappa}(\kappa_{\hat{\theta}}(x), \kappa)] \quad \leq \inf_{h_\theta \in \mathcal{H}} K\mathbb{E}[W_1(h_\theta(x), y)] + 32K^2 C_M \mathcal{R}_N(\mathcal{H}^0) + 2KC_M\sqrt{\frac{\log(1/\delta)}{2N}} \tag{8}$$

The proof involves obtaining an ERM bound using $\hat{\mathcal{R}}_N(\mathcal{L})$ and associating $\hat{\mathcal{R}}_N(\mathcal{L})$ to $\hat{\mathcal{R}}_N(\mathcal{H}^0)$ using $\hat{\mathcal{R}}_N(\mathcal{L}) \leq 8KC_M\hat{\mathcal{R}}_N(\mathcal{H}^0)$ where $\hat{\mathcal{R}}_N$ is the empirical/sample Rademacher complexity.

4 Convergence of Finite and Infinite Mixture Models

Mixture models are a very popular class of unsupervised machine learning algorithms. Often, practitioners fit a mixture model and try to interpret the cluster parameters as being representative of population cohorts. However, it is not clear if such interpretations are justified. It might be possible to fit a mixture model with different model parameters than the latent ground truth and still be able to approximate the mixture density very well. [2] establishes the conditions for the convergence of mixing distributions and provides posterior contraction rates for finite and infinite mixture models under smooth and supersmooth likelihoods. Works prior to [2] either focused on the convergence of the posterior distribution of the data density p_G or studied the convergence of cluster parameters for univariate and finite mixture models [26, 27].

If a sequence of discrete probability measures G_n with k distinct atoms converges to G_0 in the r-Wasserstein metric, then the atoms of G_n must also converge to those of G_0 after some permutation of atom labels. As a result, studying the convergence of mixing distributions of mixture models in the Wasserstein space is an intuitive way of understanding the identifiability of these models and establishing corresponding rates of convergence.

Consider a discrete probability measure $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$. It can be combined with a likelihood density $f(\cdot|\theta)$ to yield a mixture density: $p_G(x) = \int f(x|\theta) dG(\theta) = \sum_{i=1}^{k} p_i f(x|\theta_i)$. Similar to G, consider $G' = \sum_{i=1}^{k} p'_i \delta_{\theta'_i}$. The OT distance between G and G' on an underlying ground metric (Θ, ρ_ϕ) is given as

$$d_{\rho_\phi}(G, G') \quad = \inf_{q \in Q(p,p')} \sum_{i,j} q_{ij} \rho_\phi(\theta_i, \theta'_j) \tag{9}$$

Here $\theta, \theta' \in \Theta$. Also, ϕ is a convex function which induces an f-divergence between probability densities: $\rho_\phi(f_i, f'_j) = \int \phi(f'_j/f_i) f_i d\mu$ Examples include the squared Hellinger distance for $\phi(u) = \frac{1}{2}(\sqrt{u} - 1)^2$, total variation distance for $\phi(u) = \frac{1}{2}|u - 1|$, and KL divergence for $\phi(u) = -\log(u)$. Each ϕ corresponds to a particular f-divergence ρ_ϕ, which induces a composite transportation distance d_{ρ_ϕ}.

Lemma 4. *Let $G, G' \in \mathcal{G}(\Theta)$ such that both $\rho_\phi(p_G, p_{G'})$ and $d_{\rho_\phi}(G, G')$ are finite for some convex function ϕ. Then, $\rho_\phi(p_G, p_{G'}) \leq d_{\rho_\phi}(G, G')$.*

The above lemma illustrates that the f-divergence between mixture distributions is dominated by the composite transportation distance between the mixing measures G and G'. In this sense, d_{ρ_ϕ} yields a stronger topology on $\mathcal{G}(\Theta)$ than the corresponding f-divergence ρ_ϕ on the mixture densities p_G. Convergence of mixture densities may not necessarily imply convergence of the underlying discrete mixing distribution which is studied by [2].

The paper establishes Wasserstein metric identifiability of model parameters for finite mixture models, infinite convolution mixture models, and infinite mixture models with Dirichlet process prior on the clusters. Various proofs require the likelihood function to be either finite identifiable or strongly identifiable, the latter being stronger of the two condition. The paper also derives the posterior contraction rates for two types of mixture models: finite mixtures of multivariate distributions, and infinite Dirichlet process mixtures.

Theorem 5. *Required assumptions:*

(A1) The underlying space Θ is compact, and the likelihood functions $f(\cdot|\theta)$ are strongly identifiable.

(A2) $K(f_i, f_j') \leq C_1||\theta_i - \theta_j'||^2$ for any $\theta_i, \theta_j' \in \Theta$.

(A3) For any $Ginsupport(\Pi)$, $\int p_{G_0}(\log(\frac{p_{G_0}}{p_G}))^2 < C_2 K(p_{G_0}, p_G)$

(A4) Under prior Π, for small positive δ, $c_3\delta^k \leq \Pi(|p_i - p_i^*| \leq \delta) \leq C_3\delta^k$ and $c_3\delta^{kd} \leq \Pi(||\theta_i - \theta_i^*| \leq \delta) \leq C_3\delta^{kd}$

(A5) Under prior Π, all p_i as well as all pairwise distances $||\theta_i, \theta_j||$ are bounded away from 0.

Under the above assumptions, the posterior distribution of G contracts to the groundtruth G_0 under the L_2 Wasserstein distance metric at a rate of $\frac{(\log n)^{1/4}}{n^{1/4}}$.[1]

The proof uses another theorem which under certain conditions establishes that $\Pi(G|W_2(G_0, G) \geq M_n\epsilon_n|X_1, ..., X_n) \to 0$ in P_{G_0} probability. Here, ϵ_n is a sequence such that $n\epsilon_n$ is bounded away from 0 or tends to infinity. and M_n is a corresponding sequence assumed to satisfy certain intricate conditions involving packing numbers in the space $\bar{\mathcal{G}}(\Theta)$ of all discrete measures including those with countably infinite support. Taking ϵ_n to be a sufficiently large multiple of $(\log n/n)^{1/2}$ and M_n to be a large multiple of $\epsilon_n^{-1/2}$ is shown to satisfy all required conditions, thereby providing the posterior contraction rate $\frac{(\log n)^{1/4}}{n^{1/4}}$ which is minimax optimal upto a logarithmic rate for univariate finite mixtures as proved by [26].

Theorem 6. *Required assumptions:*

(A1) The Lebesgue density of the base measure P_0 is bounded away from zero. Also, it places full support on a bounded set $\Theta \subset \mathbb{R}^d$.

(A2) $K(f_i, f_j') \leq C_1\rho^{m_1}(\theta_i, \theta_j')$ for any $\theta_i, \theta_j' \in \Theta$.

(A3) For any $G \in$ support(Π), $\int p_{G_0}(\log(\frac{p_{G_0}}{p_G}))^2 < C_2 K(p_{G_0}, p_G)^{m_2}$

Under the above assumptions, there is a sequence $\beta_n \to 0$ such that $\Pi(W_2(G_0, G) \geq \beta_n|X_1, ..., X_n) \to 0$ in P_{G_0} probability. For ordinary smooth likelihood functions like the Laplacian density, the posterior contraction rate of G is dictated by $\beta_n \asymp (\log n/n)^{2/((d+2)(4+(2\beta+1)d'))}$ For supersmooth likelihood functions such as the Gaussian density, $\beta_n \asymp (\log n)^{-1/\beta}$

The proof proceeds by considering a sequence ϵ_n as a large multiple of $(\log n/n)^{1/(d+2)}$. A corresponding sequence $M_n = R_{\bar{\mathcal{G}}(\Theta)}(8\epsilon_n^2(C+4))/\epsilon_n$ is constructed where $R_{\mathcal{G}}(t)$ is defined as the inverse of Hellinger information function of the W_2 metric on the space $\bar{\mathcal{G}}$. Hence $\beta_n = M_n\epsilon_n = R_{\bar{\mathcal{G}}(\Theta)}(8\epsilon_n^2(C+4))$. Under ordinary smoothness such as that of the Laplacian density, $R_{\bar{\mathcal{G}}(\Theta)}(t) = t^{1/(4+(2\beta+1)d+\delta)}$ for some positive δ. Hence $\beta_n \asymp (\log n/n)^{2/((d+2)(4+(2\beta+1)d'))}$. For supersmooth densities such as the Gaussian density, $R_{\bar{\mathcal{G}}(\Theta)}(t) = (1/\log(1/t))^{1/\beta}$ yielding $\beta_n \asymp (\log(1/\epsilon_n))^{-1/\beta} \asymp (\log n)^{-1/\beta}$

5 Domain Adaptation with Optimal Transport

Optimal transport is also used widely in domain adaptation [17]. In domain adaptation, one has access to labeled examples from a source domain and unlabeled examples from a target domain. The goal is to predict labels for examples from the target domain. This is different from the typical train-test paradigm in machine learning, because covariate shift between the two domains is allowed.

[1] There seems to be a typo in stating this result in the paper's *Theorem 5* where it is stated as $\frac{(\log n)^{1/2}}{n^{1/4}}$ whereas the paper introduction and proof correctly state and derive it as $\frac{(\log n)^{1/4}}{n^{1/4}}$.

As a solution, optimal transport is used to map examples from the target domain to the source domain using the marginal probabilities of examples in the two domains. The problem can be easily extended to multiple source domains, each with its own covariate structure which can differ between source domains.

Theorem 7. *Given two samples X_S and X_T of sizes N_S and N_T drawn i.i.d. from source and target domains respectively, $\hat{\mu}_S$ and $\hat{\mu}_T$ being the empirical probabilities with dirac-delta masses at the observed datapoints, any $d' > d$ and $\psi' < \sqrt{2}$, there exists $N_0(d')$ such that for any $\delta > 0$ and $\min(N_S, N_T) \geq N_0(d') \max(\delta^{-(d'+2)}, 1)$ with probability at least $1 - \delta$ for all h, the following statistical bound holds:*

$$\mathcal{R}_T(h) \quad \leq \quad \mathcal{R}_S(h) + W_1(\hat{\mu}_S, \hat{\mu}_T) + \sqrt{\frac{2}{\psi'} \log\left(\frac{1}{\delta}\right)} \left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \lambda \tag{10}$$

where λ is the minimal combined error of the ideal hypothesis h^* that minimizes the combined error $\mathcal{R}_S(h) + \mathcal{R}_T(h)$.

The above statistical bound bears a striking resemblance to a similar bound for domain adaptation obtained using $\mathcal{H}$-divergence instead of Wasserstein distance [28]. The above bound between source and target risks is then used in [4] to obtain a bound between the empirical risk in the target domain $\hat{\mathcal{R}}_T(h)$ and the risk of the optimal target hypothesis $\mathcal{R}_T(h_T^*)$ as given in the following theorem:

Theorem 8. *Let D be a labeled dataset of size n. Here βn points belong to the target domain and $(1 - \beta)n$ points belong to the source domain, and $\beta \; in (0, 1)$. If $\hat{h}$ is the empirical minimizer of $\hat{\mathcal{R}}_\alpha(h) = \alpha \hat{\mathcal{R}}_T(h) + (1 - \alpha)\hat{\mathcal{R}}_S(h)$ and $h_T^* = \min_h \mathcal{R}_T(h)$, then for any $\delta \in (0, 1)$, the following is true with probability at least $1 - \delta$ over the choice of samples:*

$$\mathcal{R}_T(\hat{h}) \quad \leq \quad \mathcal{R}_T(h_T^*) + c_1 + 2(1 - \alpha) W_1(\hat{\mu}_S, \hat{\mu}_T) + \lambda + c_2 \tag{11}$$

where

$$c_1 = 2\sqrt{\frac{2K\left(\frac{(1-\alpha)^2}{(1-\beta)} + \frac{\alpha^2}{\beta}\right)\log(\frac{2}{\delta})}{n}} + 4\sqrt{\frac{K}{n}}\left(\frac{\alpha}{n\beta\sqrt{\beta}} + \frac{(1-\alpha)}{n(1-\beta)\sqrt{1-\beta}}\right)$$

and

$$c_2 = \sqrt{\frac{2}{\psi'} \log\left(\frac{1}{\delta}\right)} \left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right)$$

Another similar bound for domain adaptation is due to [3] which assumes a loss function $\mathcal{L}$ that is bounded, symmetric, k-lipschitz, and satisfies the triangle inequality. Optimal transport is used to obtain the optimal coupling γ^* between the source data and target data:

$$\gamma^* = \underset{\gamma \in \Pi(P_s, P_t)}{\text{argmin}} \int (\alpha d(x_s, x_t) + \mathcal{L}(y_s, y_t)) d\gamma(x_s, y_s, x_t, y_t)$$

The labeling function $h \in \mathcal{H}$ is bounded i.e. $|h^*(x_1) - h^*(x_2)| \leq M$. Again, N_S and N_T denote number of source and target datapoints. Then for all $\lambda > 0$ and $\alpha = k\lambda$, we have with probability at least $(1 - \delta)$ that:

$$\mathcal{R}_T(h) \quad \leq \quad + W_1(\hat{\mu}_S, \hat{\mu}_T) + \sqrt{\frac{2}{\psi'} \log\left(\frac{2}{\delta}\right)} \left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \lambda + kM\phi(\lambda) \tag{12}$$

where λ is again the minimal combined error of the ideal hypothesis h^* that minimizes the combined error $\mathcal{R}_S(h) + \mathcal{R}_T(h)$.

Thus, [4, 3] provide a theoretical basis for why optimal transport can be used for domain adaptation.

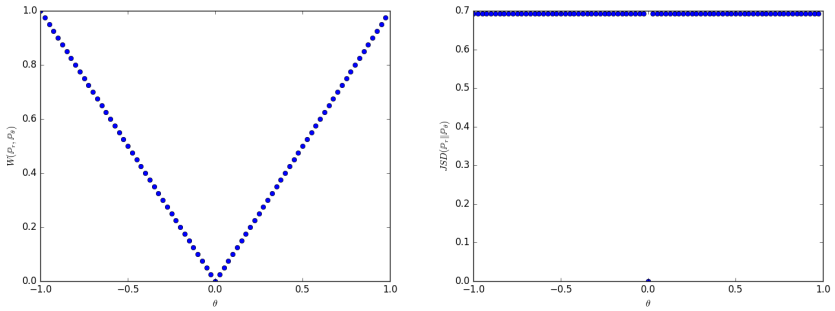

Figure 1: Figure from the WGAN paper [5] showing the utility of gradients provided by Wasserstein distance loss versus Jensen-Shannon divergence loss. Wasserstein loss (left) is continuous and provides a useful gradient, whereas the JS divergence loss (right) is discontinuous and the gradient does not seem very useful for learning in the space of probability distributions.

6 Wasserstein Generative Adversarial Networks

Another popular model that makes use of optimal transport is Wasserstein Generative Adversarial Network (WGAN) [5] . WGAN is a member of the family of models called Generative Adversarial Networks (GANs) [29]. In a GAN, there are two predictive models, both of which are typically deep neural networks. One of the networks called the generator network generates samples in the domain of the dataset that are as realistic as possible. The discriminator network tries to distinguish between the real samples in the dataset and the fake samples generated by the generator network. Thus, the training of GANs can be viewed as a game between the two networks where generator network which tries to fool the discriminator network and the discriminator network tries to accurately tell apart the generator output from real data.

WGAN improves upon previous GAN losses and provides a more stable training procedure that is less prone to mode collapse i.e. abrupt training failure. Original GAN used the Jensen-Shannon divergence for training which is not continuous everywhere and differentiable almost everywhere. Other losses such as KL-divergence that have been used with GANs have similar issues. However, Wasserstein loss based on 1-Wasserstein distance proposed in [5] is continuous everywhere and differentiable almost everywhere, which provides a training signal in all regions of the feature space and prevents abrupt mode collapses that happened with previous GAN losses.

A primary assumption that is used in the theoretical analysis of WGAN if the finiteness of the mean of the local Lipschitz constant. If the generator neural network is denoted as $g_\theta(z)$ where θ are parameters of the neural network and z is input to the generator from some distribution $z \sim Z$ and g is locally Lipschitz with the local Lipschitz constants $L(\theta, z)$, then g needs to satisfy $\mathbb{E}_z[L(\theta, z)] < +\infty$. The following theorems assume this property of g.

Theorem 9. *If g_θ is a feedforward neural network parametrized by θ (and therefore as a function continuous in θ) and $p(z)$ is a prior over $z \sim Z$ such that $\mathbb{E}_z[||z||] < \infty$, $\mathbb{P}_r$ is the real data distribution over $\mathcal{X}$, and $\mathbb{P}_\theta$ is the distribution of $g_\theta(\mathcal{Z})$, then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere and differentiable almost everywhere.*

The paper provides an example to show that other distances or divergences such as Jensen-Shannon divergence or Kullback-Leibler divergence do not enjoy this nice property of the Wasserstein distance and hence lead to instability while training the generator. Since the OT computation is highly intractable in continuous function spaces, [5] uses the dual problem as follows:

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{||f||_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(Z)}[f_w(g_\theta(z))] \tag{13}$$

If the constraint $||f||_L \leq 1$ is removed, we obtain Wasserstein distances upto a multiplicative constant which is good enough for training, because we are not interested in the computation of the exact Wasserstein distance, but only in its use as an appropriate loss function for training a GAN.

Theorem 10. *Given the real data distribution $\mathbb{P}_r$ and the generative data distribution $\mathbb{P}_\theta$ based on the distribution $p(Z)$ over $\mathcal{Z}$ and the generator feedforward neural network $g_\theta(z)$ which satisfies the*

assumption $\mathbb{E}_z[L(\theta, z)] < +\infty$ described before, there exists a function $f : \mathcal{X} \to \mathbb{R}$ to the problem 13, and the gradient of the Wasserstein distance is given as

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = -\mathbb{E}_{z\sim p(z)}[\nabla_\theta f(g_\theta(z))] \tag{14}$$

The gradient of the Wasserstein loss can now be used to backpropagate through equation 13 by estimating $\mathbb{E}_{z\sim p(z)}[\nabla_\theta f(g_\theta(z))]$.

7 Selected Applications

Wasserstein distance is remarkably rich due to its ability to take the underlying geometry of a measure's domain into account. This allows it to naturally incorporate domain information into the machine learning solution through the specification of an appropriate ground metric. Though the specification of ground truth is an additional requirement of Wasserstein metric, it is often available naturally e.g. word embeddings in natural language, object descriptors in computer vision, etc. This distinguishing characteristic of Wasserstein distance has been used in specific domains such as natural language processing, computer vision, etc. to build state-of-the-art methods for machine learning tasks.

As an example, we present the application of optimal transport to calculating distances between documents. [12] presents an unsupervised method for calculating *Word Mover's Distance* by transporting the histogram of words between two documents and using the Euclidean distance between word embeddings as the ground metric. [13] further proposed *Supervised Word Mover's Distance* by learning an affine transformation of word embeddings and a word-importance weight vector using label supervision for minimizing the stochastic LOO nearest neighbor classification error. This application was the author's introduction to Wasserstein distances and optimal transport. We have provided a representative list of applications of optimal transport in machine learning in the introductory section of this report.

8 Conclusion

In this report, we surveyed recent theoretical results underpinning the use of optimal transport in statistical machine learning. In particular, we focused on the use of Wasserstein loss on training multi-class and multi-label classifiers, and generative adversarial networks. We also examined the risk bounds for the case when optimal transport is used for domain adaptation. Finally, we examined the convergence and posterior contraction rates established recently for finite and infinite mixture models using composite Wasserstein distance.

8.1 Potential Open Problems

A literature survey like this is an opportunity to identify potential avenues for future research. Some of these are identified below.

- **Robust Optimal Transport**: Can we use a more robust specification of the underlying ground metric that specifies not just the cost but also the uncertainty associated with it?
- **Extreme Classification with Wasserstein Loss**: Wasserstein loss provides a way to incorporate side information about the ground metric between labels into the classification problem. Can training with the Wasserstein loss be scaled up to the regime of extreme classification where such side information would be extremely valuable in making good predictions?
- **Connections with Kernel/Metric Learning**: The specification of a ground cost immediately associates the Wasserstein metric with a corresponding kernel. Can this ground metric be learned for challenging structured spaces such those of trees, graphs, strings, etc.? There is associated recent work on ground metric learning [19] and sliced Wasserstein kernels [18].

We hope to further understand the connections between the Wasserstein metric and its role in capturing geometric information about machine learning tasks, and use it in solving our machine learning problems.

References

[1] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[2] XuanLong Nguyen et al. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[3] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3733–3742, 2017.

[4] 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*, pages 737–753. Springer, 2017.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[6] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[7] L Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, pages 227–229, 1942.

[8] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[9] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. Technical report, 2017.

[10] Sigurd Angenent, Steven Haker, and Allen Tannenbaum. Minimizing flows for the monge–kantorovich problem. *SIAM journal on mathematical analysis*, 35(1):61–97, 2003.

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[12] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[13] Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, pages 4862–4870, 2016.

[14] Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[15] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[16] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport.

[17] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

[18] Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

[19] Marco Cuturi and David Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

[20] Peng Zhao and Zhi-Hua Zhou. Label distribution learning by optimal transport. 2018.

[21] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. In *International Conference on Learning Representations*, 2018.

[22] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[23] Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pages 2492–2500, 2012.

[24] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems*, pages 4197–4205, 2016.

[25] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[26] Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, pages 221–233, 1995.

[27] Hemant Ishwaran, Lancelot F James, and Jiayang Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456):1316–1332, 2001.

[28] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.