Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

A CLASS OF OPTIMAL TRANSPORT REGULARIZED FORMULATIONS WITH APPLICATIONS TO WASSERSTEIN GANS

Saied Mahdian
Jose H. Blanchet
Peter W. Glynn

Department of Management Science and Engineering
Stanford University
Stanford, CA 94305, USA

ABSTRACT

Optimal transport costs (e.g. Wasserstein distances) are used for fitting high-dimensional distributions. For example, popular artificial intelligence algorithms such as Wasserstein Generative Adversarial Networks (WGANs) can be interpreted as fitting a black-box simulator of structured data with certain features (e.g. images) using the Wasserstein distance. We propose a regularization of optimal transport costs and study its computational and duality properties. We obtain running time improvements for fitting WGANs with no deterioration in testing performance, relative to current benchmarks. We also derive finite sample bounds for the empirical Wasserstein distance from our regularization.

1 INTRODUCTION

Optimal transport costs, which include the Wasserstein distance and the earth-mover distance as special cases, have become useful tools in machine learning and statistics (Kolouri et al. 2017; Arjovsky et al. 2017; Abadeh et al. 2015; Kusner et al. 2015; Cuturi 2013; Blanchet and Murthy 2019). The Wasserstein distance is a powerful statistical tool which can be used to compare arbitrary probability distributions defined on general spaces involving complex geometrical properties and high-dimensional features (see, for example, (Villani 2003)). So, for example, one can use the Wasserstein distance to compare discrete vs continuous distributions directly, without introducing smoothing, in contrast to alternatives such as the Kullback-Leibler divergence (see (Arjovsky et al. 2017; Genevay et al. 2016) for more details). Also, by judiciously choosing the underlying metric, a Wasserstein distance can generate either the topology corresponding to weak convergence or the total variation distance.

The optimal transport cost between distribution μ and ν with respect to the (non-negative) transportation cost function $\widetilde{c}(\cdot)$, denoted as $D_{\widetilde{c}}(\mu,\nu)$, is computed as the solution of a linear programming (LP) problem. Precisely, suppose that X follows distribution μ, while Y follows distribution ν. Then, $D_{\widetilde{c}}(\mu,\nu)$ is obtained by minimizing $E_\pi[\widetilde{c}(X,Y)]$ (the expected transportation cost) over all joint distributions π of the pair (X,Y) satisfying the marginal distributional constraints that $X \sim \mu$ and $Y \sim \nu$. (If $\widetilde{c}$ is a metric then $D_{\widetilde{c}}(\mu,\nu)$ is the Wasserstein distance with respect to $\widetilde{c}$ and we write $D_{\widetilde{c}}(\mu,\nu) = \mathscr{W}_{\widetilde{c}}(\mu,\nu)$ – we often omit the subscript $\widetilde{c}$ unless there is the potential for confusion.)

While there are algorithms that enable the evaluation of the Wasserstein distance in many settings, solving these types of LPs inside iterative routines, for example for fitting parametric families via the minimization of the empirical Wasserstein distance, remains challenging.

Our main contribution in this paper involves introducing regularization tools which facilitate the types of fitting procedures involving the minimization of the Wasserstein distance over a parametric family of distributions. In addition, owing to these regularization tools, we obtain finite sample bounds for the empirical Wasserstein distance.

An application that is interesting involves so-called WGAN (Wasserstein Generative Adversarial Network) simulators, which are of significant interest in artificial intelligence, see (Arjovsky et al. 2017).

A WGAN simulator is typically calibrated as follows. Suppose that μ_n encodes a target empirical measure which is a proxy for μ_∞ and from which we wish to generate objects. Assume that $\{\mu(\theta) : \theta \in \Theta\}$ is a parametric family of distributions that we wish to use to approximate μ_∞. We are interested in solving

$$\min_{\theta\in\Theta} \mathscr{W}_{\widetilde{c}}(\mu(\theta), \mu_n) = \min_{\theta\in\Theta} \sup_{f:|f(x)-f(y)|\leq \widetilde{c}(x,y)} E_{\mu(\theta)}[f(X)] - E_{\mu_n}[f(Y)]. \tag{1}$$

The functions f in the inner sup are known as 'actor critics'. These functions are typically parameterized using a family of neural networks. Moreover, $\mu(\theta)$ is often encoded as the output of another neural network family with random (e.g. Gaussian) initial input and with parameters encoded by θ. The WGAN can be interpreted via (1) as a procedure in which two different neural networks are interacting against each other.

In order to calibrate $\mu(\theta)$ it is common to use an iterative procedure which involves an inner and an outer loop corresponding to the min-max structure of the WGAN. The inner loop involves the evaluation of an optimal actor critic function f (which depends on the current parameter θ to be updated at the outer loop). A fair amount of experimentation and implementation tricks have been developed which can yield convergence to a reasonable solution, (Gulrajani et al. 2017). All of these tricks can be easily adapted to our regularization formulation and we are able to show improved training time in our numerical experiments.

WGANs represent just one of many data-driven settings in which Wasserstein distance can be used as a fitting tool. In all of them, the distribution μ_n is used as a surrogate for an underlying measure μ_∞ which is the target that one wishes to learn. It is natural then to recognize that μ_n is an imperfect/noisy description of μ_∞ and that any other distribution which is reasonably close to μ_n (in the sense of being indistinguishable given the statistical noise associated to any finite sample) should yield a similar performance to that of μ_n. This perspective is particularly important in light of the fact that non-parametric empirical estimators of the Wasserstein distance converge slowly (at rate $O\left(n^{-1/d}\right)$ where d is the underlying dimension of the distribution and n is the number of samples, see (Dudley 1969; Weed and Bach 2019). So, it is natural to take the view that plausible variations of the data can be used to facilitate the estimation of optimal transport costs.

Using this insight, we provide a formulation which regularizes $D_{\widetilde{c}}(\mu_n, \mu)$. In particular, our regularization formulation takes the generic form

$$G_\delta(\mu_n, \mu) = \inf\{D_{\widetilde{c}}(\nu, \mu) : \nu \in \mathscr{D}_\delta(\mu_n)\}, \tag{2}$$

where $\mathscr{D}_\delta(\mu_n) = \{\nu : D_c(\mu_n, \nu) \leq \delta\}$, for some optimal transport cost $D_c(\mu_n, \nu)$ depending on a cost function c. As $\delta \to 0$, under mild continuity assumptions, we recover the standard optimal transport cost.

The map $\mu_n \mapsto G_\delta(\mu_n, \mu)$ is intuitively a more regular object than $G_0(\mu_n, \mu) = D_{\widetilde{c}}(\mu_n, \mu)$ as it is less sensitive to small perturbations of μ_n. Of course, this type of regularity is also achieved by maximizing over a neighborhood of μ_n (instead of minimizing), but this operation leads to computational complications because the optimal transport cost is a convex functional. The dual formulation of the optimal transport costs can be used to connect our regularization, at least formally, to smoothing techniques that are often used in the non-smooth convex optimization literature (Nesterov 2005).

As indicated earlier, the regularization approach that we take is particularly meaningful given the slow rates of convergence in the empirical estimation of Wasserstein distances. Moreover, since the estimated Wasserstein distance is a positive random variable, the statistical error is likely to often have a right-tail bias, thus the minimization operation that we apply in (2) to regularize the Wasserstein distance is also sensible as a means of mitigating this bias. However, we need to be careful to not overcompensate. So, we also provide statistical learning bounds which can be used to ensure a choice of δ which enables the use of $G_\delta(\mu_n, \mu)$, plus a small correction term, as an upper bound for $D_{\widetilde{c}}(\mu_\infty, \mu)$. These statistical learning bounds are presented in Theorem 4. The parameter $\delta > 0$ could also be chosen by a cross-validation procedure.

There are other regularization methods for estimating optimal transport costs. Some of these techniques require some smoothness or absolute continuity between the measures involved; this occurs, for example,

when using entropic regularization, (Cuturi 2013; Sanjabi et al. 2018). Others impose low rank constraints, as in (Forrow et al. 2019), in the setting of domain adaptation, and others (Sanjabi et al. 2018; Gulrajani et al. 2017; Gao et al. 2017) focus on specific applications such as Wasserstein GANs. In particular, (Gao et al. 2017) study an empirical risk minimization framework and propose a similar WGAN formulation to ours but our formulation does not require cost functions to be differentiable, we provide expressions which are simpler to use for computation, and we also supply out-of-sample generalization bounds.

Our regularization technique does not require smoothing or low rank properties. It acts directly at the same level of generality as the original optimal transport formulation. However, we are able to show that $G_\delta(\mu_n,\mu)$ can often be evaluated directly and conveniently in terms of $D_{\widetilde{c}}(\mu_n,\mu)$, leading to a variation of the optimal transport cost formulation which can then be used in conjunction with any of the regularization methods mentioned earlier. So, we do not see our work as a competitor to these regularization methods. Our approach can be reasonably viewed a pre-conditioning step which can be applied before any regularization tool that uses additional data structure.

In the context of WGANs, in Section 2 we show that under mild assumptions,

$$\min_\theta \min_{\mathscr{W}(\mu_n,\nu)\le\delta} \mathscr{W}(\mu_\theta,\nu) = \min_\theta \sup_{f:|f(x)-f(y)|\le\widetilde{c}(x,y)} \left(E_{\mu(\theta)}[f(X)] - E_{\mu_n}[f(Y)] - \delta\right)^+. \tag{3}$$

So, our regularization technique corresponds to 'flattening' of the optimization surface in the parameter space θ. The amount of flattening is governed by δ, which should correspond to the degree of ambiguity in the data, measured from a statistical point of view. This flattening has the effect of reducing the frequency of iterates of the generative network, parameterized by θ, relative to the actor critic iterates, represented by f. While this implementation device (i.e. iterating the actor critic more often than the generator) is used in practice to speed up training times, our approach is theoretically supported from an optimality perspective.

In summary, our Optimal Transport regularization (OTR) formulation suggests that training of the generative network can be reduced without loss of performance if $\delta > 0$ is well calibrated to reflect the size of plausible statistical noise, using cross validation. We validate our findings by experimenting on the image data sets MNIST and CIFAR10.

The rest of the paper is organized as follows. In Section 2 we introduce the standard optimal transport problem and present our OTR regularization formulation, convenient simplified expressions (such as (3)) and strong duality results. In Section 3, we discuss statistical learning bounds for our empirical estimator. Our numerical examples (in Section 4) suggest that our OTR estimator is often a better upper bound than the standard Wasserstein estimator. The proofs of all theorems are provided in the Appendix.

2 PROBLEM FORMULATION, INTERPRETATIONS AND TRACTABILITY

We start by formulating the standard optimal transport problem. To do so, we shall introduce notation which will also be useful when describing our proposed formulation. Throughout the paper we will consider distributions supported on metric spaces $\mathscr{S}_X$ and $\mathscr{S}_Y$ with metrics d_X and d_Y, respectively. We assume, for simplicity in the exposition that the spaces are complete, separable and compact.

We shall use X to denote a generic random variables taking values in $\mathscr{S}_X$. Likewise, a generic random variable Y will take values in $\mathscr{S}_Y$. The space of Borel probability measures defined on $\mathscr{S}_X$ and $\mathscr{S}_Y$ are defined as Π_X and Π_Y, respectively. We use $\Pi_{X,Y}$ to denote the set of all couplings between X,Y (i.e. joint Borel probability measures on $\mathscr{S}_X\times\mathscr{S}_Y$). Further, $\Pi_{X,Y}(\mu_0,\nu)$ is the subset of $\Pi_{X,Y}$ such that $X\sim\mu_0$ and $Y\sim\nu$ (i.e. X follows distribution μ and Y follows distribution ν).

Given a generic element $\pi\in\Pi_{X,Y}$, π_X is the marginal distribution of X and π_Y is the marginal distribution of Y. So, $\pi\in\Pi_{X,Y}(\mu_0,\nu)$ implies that $\pi_X=\mu_0$ and $\pi_Y=\nu$.

The standard optimal transport problem, also known as the Monge-Kantorovich problem, can be written as (see (Villani 2003))

$$\mathscr{P}_0 : D_{\widetilde{c}}(\mu_0,\nu) = \min\{\mathbb{E}_\pi\widetilde{c}(X,Y) : \pi\in\Pi_{X,Y}(\mu_0,\nu)\}$$

where $\widetilde{c}: \mathscr{S}_X \times \mathscr{S}_Y \to [0,\infty)$ is a lower semicontinuous function. Clearly, $D_{\widetilde{c}}(\mu_0,\nu)$ is the solution of a linear programming problem (albeit, an infinite dimensional one). We now consider the corresponding dual. First, let $C(\mathscr{S}_X)$ and $C(\mathscr{S}_Y)$ be the space of continuous functions on $\mathscr{S}_X$ and $\mathscr{S}_Y$, respectively. Next, define $\mathscr{A}(\widetilde{c}) = \{(\alpha,\beta) \in C(\mathscr{S}_X) \times C(\mathscr{S}_Y) : \alpha(x)+\beta(y) \leq \widetilde{c}(x,y) \text{ for all } x \in \mathscr{S}_X, y \in \mathscr{S}_Y\}$, then, the dual problem formulation of $\mathscr{P}_0$ is

$$\bar{\mathscr{P}}_0 : \sup\{\mathbb{E}_{\mu_0}\alpha(X) + \mathbb{E}_{\nu}\beta(Y) : (\alpha,\beta) \in \mathscr{A}(\widetilde{c})\}.$$

It is known (see (Villani 2003)) that strong duality holds.

To define our relaxed optimal transport formulation, we introduce the region $\mathscr{D}_\delta(\mu_0) = \{\nu : D_c(\mu_0,\nu) \leq \delta\}$. We employ a lower semicontinuous cost function $c: \mathscr{S}_X \times \mathscr{S}_X \to [0,\infty)$ satisfying $c(x,x) = 0$, so that $\mathscr{D}_0(\mu_0) = \{\mu_0\}$. As indicated in (2), we are interested in $G_\delta(\mu_0,\nu) = \min\{D_{\widetilde{c}}(\mu,\nu) : \mu \in \mathscr{D}_\delta(\mu_0)\}$. We have replaced the inf in (2) by min because $\mathscr{D}_\delta(\mu_0)$ is a compact set in the weak convergence topology (Prohorov's theorem) and the optimal transport cost, as the supremum of linear and continuous functionals (by duality), is lower semicontinuous.

In terms of the dual problem $\bar{\mathscr{P}}_0$, $G_\delta(\mu_0,\nu) = \min_{\mu\in\mathscr{D}_\delta(\mu_0)} \sup_{(\alpha,\beta)\in\mathscr{A}(\widetilde{c})} \mathbb{E}_\mu\alpha(X) + \mathbb{E}_\nu\beta(Y)$ is the form our relaxed formulation takes. The next result indicates that duality holds in this representation, meaning, that min and sup can be exchanged, this will serve to provide useful interpretations for $G_\delta(\mu_0,\nu)$.

Theorem 1 $G_\delta(\mu_0,\nu) = \sup_{(\alpha,\beta)\in\mathscr{A}(\widetilde{c})} \min_{\mu\in\mathscr{D}_\delta(\mu_0)} \mathbb{E}_\mu\alpha(X) + \mathbb{E}_\nu\beta(Y)$.

The above theorem can be used to provide a formal interpretation of our regularization as a smoothing technique related to Nesterov's smoothing (Nesterov 2005). We have

$$G_\delta(\mu_0,\nu) = \sup_{-\alpha,\beta\in\mathscr{A}(\widetilde{c})} \inf_{\mu\in\mathscr{D}_\delta(\mu_0)} E_\nu\beta(Y) - E_\mu\alpha(X) = \sup_{-\alpha,\beta\in\mathscr{A}(\widetilde{c})} (E_\nu\beta(Y) - \phi(\alpha;\mu_0)) \tag{4}$$

where $\phi(\alpha;\mu_0) = \sup_{\mu\in\mathscr{D}_\delta(\mu_0)} E_\mu\alpha(X)$ is a convex function of α. The above representation coincides in form with the smoothing operator technique introduced by Nesterov, see (Nesterov 2005), equation (2.2). The resulting smooth mapping in Nesterov's representation is to be considered as a function of ν, namely $\nu \mapsto G_\delta(\mu_0,\nu)$. While we believe that it is interesting to study the transformation (4) in future research for the purpose of smoothing optimal transport problems, we shall focus on studying $G_\delta(\mu_0,\nu)$. Note that controlling the size of δ will guarantee the validity of statistical bounds when estimating optimal transport costs from empirical data.

In addition to the smoothing interpretation given by (4), Theorem 1 also admits an economic interpretation. Consider an agent who offers a transportation service to two customers. One of them wishes to transport a pile of sand out of his/her backyard (this pile of sand is modeled according to distribution μ_0), while the other customer wishes to cover a sinkhole in his/her own backyard (the profile of the sinkhole is modeled by distribution ν). It would cost $c(x,y)$ to transport mass from location x to location y if the customers arrange to solve this transportation problem among themselves. So, the agent would wish to charge a price $\alpha(x)$ per unit of mass transported from location x to the first customer, a price $\beta(y)$ per unit of mass transported from location y to the second customer, and would do so in such a way that it is cheaper to pay these prices than to pay the cost of transporting directly without the intervention of the agent, so $\alpha(x)+\beta(y) \leq c(x,y)$. But, of course, the agent wishes to maximize the total profit and this yields the dual interpretation for transporting items, encoded by distributions μ_0,ν. Theorem 1 indicates that $G_\delta(\mu_0,\nu)$ solves a distributionally robust revenue maximization problem, in which the agent selects a policy which is robust to perturbations in the shape of the pile of sand reported by the first customer.

Next, we provide another representation for $G_\delta(\mu_0,\nu)$, which forms the basis for the design of gradient and subgradient algorithms and further simplifications.

Theorem 2 $G_\delta(\mu_0,\nu) = (-1)\cdot\min_{\lambda\geq 0} \left\{\lambda\delta + \max_{\pi\in\Pi_{W,Y}(\mu_0,\nu)} \mathbb{E}_\pi[h(W,Y,\lambda)]\right\}$ where $h: \mathscr{S}_X \times \mathscr{S}_Y \times \mathbb{R}_+ \to \mathbb{R}$ and $h(w,y,\lambda) = \sup_x\{-\widetilde{c}(x,y) - \lambda c(x,w)\}$.

The above result provides further insight into the smoothness properties introduced by our regularization technique. For instance, min-max representation justifies understanding our regularization as a regularization technique as in (Esfahani and Kuhn 2018; Blanchet et al. 2019). Also, consider the case $\mathscr{S}_X = \mathscr{S}_Y$ and $c = d_X$. Then, the function $h(w,z,\lambda)$ becomes λ-Lipschitz in the w argument. In particular, for all $w_1, w_2 \in \mathscr{S}_X$, $h(w_1,z,\lambda) - h(w_2,z,\lambda) \leq \sup_x \{\lambda c(x,w_2) - \lambda c(x,w_1)\} \leq \lambda d_X(w_1,w_2)$. So, Theorem 2 implies that solving for $G_\delta(\mu_0,\nu)$ is equivalent to solving a standard optimal transport problem with measures μ_0, ν and a cost function that which replaces $\widetilde{c}(x,y)$ by a cost function which is λ-Lipschitz in w and λ is regularized.

In view of Theorem 2, we define $g(\lambda,\mu_0,\nu) = \lambda\delta + \max_{\pi\in\Pi_{W,Y}(\mu_0,\nu)} \mathbb{E}_\pi[h(W,Y,\lambda)]$. Thus, $G_\delta(\mu_0,\nu) = (-1)\cdot\min_{\lambda\geq 0} g(\lambda,\mu_0,\nu)$. The function $h(\cdot)$ is convex in λ and subsequently $g(\cdot)$ is a convex function of λ. Hence, $\min_{\lambda\geq 0} g(\lambda,\mu_0,\nu)$ is a convex optimization problem. Moreover, since $\lim_{\lambda\to\infty} g(\lambda,\mu_0,\nu) = \infty$, the optimal solution set for $\min_{\lambda\geq 0} g(\lambda,\mu_0,\nu)$ is bounded. Next, we provide a result which can be used as a basis for a subgradient algorithm to compute $G_\delta(\mu_0,\nu)$.

Theorem 3 If $\mathscr{S}_X, \mathscr{S}_Y$ are convex subsets of $\mathbb{R}^d$ (for $d \in \mathbb{N}$), and $\widetilde{c}, c$ are continuous, and $\widetilde{c}(\cdot,y) + \lambda c(\cdot,w)$ is a strictly convex function for $\lambda \geq 0, w$ and y, then h is differentiable in λ. Further, the left-hand and right-hand partial derivatives of $g(\lambda,\mu_0,\nu)$ with respect to λ are $\delta + \min_{\pi\in\Pi^*(\lambda)} \mathbb{E}_\pi\left[\frac{\partial}{\partial\lambda} h(W,Y,\lambda)\right]$ and $\delta + \max_{\pi\in\Pi^*(\lambda)} \mathbb{E}_\pi\left[\frac{\partial}{\partial\lambda} h(W,Y,\lambda)\right]$ respectively where $\Pi^*(\lambda)$ is set of optimal solutions to the problem $\max_\pi \quad \mathbb{E}_{\pi\in\Pi_{W,Y}(\mu_0,\nu)}[h(W,Y,\lambda)]$.

Remark 1 Theorem 3 still holds if the strict convexity condition for $\widetilde{c}(\cdot,y) + \lambda c(\cdot,w)$ is replaced with the condition that $\arg\min_{x\in\mathscr{S}_X}\{\widetilde{c}(x,y) + \lambda c(x,w)\}$ is a singleton for all $\lambda \geq 0, w$ and y.

Remark 2 Function g is differentiable at any point λ if and only if the set $\left\{\mathbb{E}_\pi\left[\frac{\partial}{\partial\lambda} h(W,Y,\lambda)\right] \mid \pi \in \Pi^*(\lambda)\right\}$ is a singleton (for more details, see Corollary 4 of (Milgrom and Segal 2002)).

Theorem 3 suggests implementing a subgradient method (Bertsekas 2015) to solve $\min_{\lambda\geq 0} g(\lambda,\mu_0,\nu)$. In particular, at each iteration t, using λ_{t-1}, we can find $\pi_{\lambda_{t-1}}$ (a member of $\Pi^*(\lambda_{t-1})$) and then λ_t. We assume we have access to an oracle to solve for $\Pi^*(\lambda)$. Developing efficient methods to solve optimal transport problems for $\Pi^*(\lambda)$ is a topic of separate interest which we will not focus in this paper. Once we arrive at the optimal solution $(\lambda^*, \pi_{\lambda^*})$ (or a reasonable approximation of the optimal solution), then an optimal mapping between y,x solving $\min_{\lambda\geq 0} g(\lambda,\mu_0,\nu)$ can be constructed as follows.

1. For each point y, map it to a new point w using π_{λ^*}.
2. Find x as the solution to the problem $\sup_x\{-\widetilde{c}(x,y) - \lambda^* c(x,w)\}$.

We conclude this section with examples in which $G_\delta(\mu_0,\nu)$ can be substantially simplified.

Example 1 (Wasserstein Distances of Order 2) Let $\widetilde{c}(x,y) = ||x-y||^2$ and $c(x,w) = ||x-w||^2$. Then, $-G_\delta(\mu_0,\nu) = \min_{\lambda\geq 0}\left\{\delta\lambda - \left(\frac{\lambda}{1+\lambda}\right)\cdot\left(\min_{\pi\in\Pi_{W,Y}(\mu_0,\nu)}\mathbb{E}_\pi||W-Y||^2\right)\right\}$. Let $H_0 = \min_{\pi\in\Pi_{W,Y}(\mu_0,\nu)}\mathbb{E}_\pi||W-Y||^2$. Then the optimal λ is $\lambda = \left(\sqrt{\frac{H_0}{\delta}} - 1\right)^+$.

Example 2 (Wasserstein distance of order 1 and WGANs) Let $\mathscr{S}_X = \mathscr{S}_Y$ with metric $d_X = d_Y = d$. For this subsection, let $\widetilde{c}(x,y) = c(x,y) = d(x,y)$ for all $x,y \in \mathscr{S}_X$. First, we claim that if $\lambda > 1$, $h(w,y,\lambda) = -d(w,y)$ and if $\lambda < 1$, $h(w,y,\lambda) = -\lambda d(w,y)$. This can be seen as follows. Let $x_{(w,y)} = \arg\max_{x\in\mathscr{S}_X}\{-d(x,y) - \lambda d(x,w)\}$. Then for $\lambda > 1$, $-d(w,y) = -d(w,y) - \lambda\cdot d(w,w) \leq -d(x_{(w,y)},y) - \lambda\cdot d(x_{(w,y)},w) \Leftrightarrow \lambda\cdot d(x_{(w,y)},w) \leq d(w,y) - d(x_{(w,y)},y) \leq d(x_{(w,y)},w) \Leftrightarrow (\lambda-1)d(x_{(w,y)},w) \leq 0 \Leftrightarrow x_{(w,y)} = w$. A similar argument holds for $\lambda < 1$. In addition, since $h(w,y,\lambda)$ is the supremum of Lipschitz functions, it is continuous in λ. So, for $\lambda = 1$, $h(w,y,\lambda) = -d(w,y)$. For $\lambda \geq 1$, the minimum of $g(\lambda,\mu_0,\nu)$ occurs at $\lambda = 1$. For $\lambda \in [0,1]$, if $G_0(\mu_0,\nu) - \delta \leq 0$, the minimum of $g(\lambda,\mu_0,\nu)$ occurs at $\lambda = 0$; otherwise, the minimum occurs at $\lambda = 1$. So,

$$G_\delta(\mu_0,\nu) = \max\{G_0(\mu_0,\nu) - \delta, 0\}. \tag{5}$$

Expression (5) can be directly applied to training Wasserstein GANs. For additional background on these types of generative networks, see (Arjovsky et al. 2017; Gulrajani et al. 2017). Wasserstein GANs involve the optimization problem $\min_\theta G_0(\mu_n, \mu(\theta))$ where μ_n is the empirical measure of a real dataset and $\mu(\theta)$ is a parametric probability measure to be constructed using a generative network.

Our *OTR for Wasserstein GANs* takes the form

$$\min_\theta G_\delta(\mu_n, \mu(\theta)) = \min_\theta \max\{G_0(\mu_n, \mu(\theta)) - \delta, 0\} = \min_\theta \max_{f \in Lip(1)} \left(E_{\mu_n} f(X) - E_{\mu(\theta)} f(X) - \delta\right)^+, \quad (6)$$

where $Lip(1)$ represents the space of 1-Lipschitz functions with respect to the metric $d(\cdot)$. Note that $\delta = 0$ recovers the problem for Wasserstein GANs. Our implementation involves just a small modification of standard Wasserstein GAN platforms. However, it is important to choose the regularization parameter δ carefully. The next section provides statistical guidance to this effect.

Solving (6) requires only a simple augmentation to any stochastic gradient descent procedure proposed for Wasserstein GANs. In particular, (5) implies $\nabla_\theta G_\delta(\mu_n, \mu(\theta)) = \nabla_\theta G_0(\mu_n, \mu(\theta)) \mathbb{1}(G_0(\mu_n, \mu(\theta)) \geq \delta))$ where $\mathbb{1}(\cdot)$ is the indicator function. So, in a stochastic gradient descent implementation, θ should be updated only when $\mathbb{1}(G_0(\mu_n, \mu(\theta)) \geq \delta))$ and the procedure will be the same as for Wasserstein GANs. Experiment results are provided in Section 4.

3 THE STATISTICS OF THE OTR PROBLEM

In the previous section we studied the optimization problem $\min_{\lambda \geq 0} g(\lambda, \mu_0, \nu)$ with respect to any distribution μ_0. In this section, we study statistical guarantees when μ_0 is given by an empirical measure μ_n of i.i.d. observations, so its canonical representation takes the form $\mu_n(dx) = n^{-1} \sum_{j=1}^n \delta_{X_i}(dx)$, with the X_i's being i.i.d. copies of some distribution μ_∞. We derive a confidence interval for $G_0(\mu_0, \nu)$ through the use of concentration inequalities. In this section, we focus on the case where $\mathscr{S}_X = \mathscr{S}_Y$ and $d_X = d_Y$. In addition, for all $x, y \in \mathscr{S}_X$, we set $c(x,y) = d^k(x,y)$ where $k \geq 1$.

Suppose $c, \widetilde{c}$ are Lipschitz functions with Lipschitz constants $L(c)$ and $L(\widetilde{c})$ respectively. As a result, $h(w,y,\lambda)$ is Lipschitz in (w,y) with Lipschitz constant $K_\lambda = O(L(c)\lambda \vee L(\widetilde{c}))$. Define

$$\varepsilon(n, \rho, \zeta, K_\lambda) = \sqrt{\frac{\log(\frac{1}{\rho})}{2n}} + 4\zeta K_\lambda + \frac{8\sqrt{2}K_\lambda}{\sqrt{n}} \int_{\zeta/4}^{4diam(\mathscr{S}_X)} \sqrt{\mathscr{N}(\mathscr{S}_X, d_X, \xi/4) \log\left(2\left\lceil \frac{2diam(\mathscr{S}_X)}{\xi} \right\rceil + 1\right)} d\xi$$

where $\mathscr{N}(\mathscr{S}_X, d_X, \xi)$ is the ξ-covering number for $(\mathscr{S}_X, d_X)$.

Theorem 4 For $k = 1, d \geq 2, \zeta > 0, \delta \geq 0, \lambda > L(\widetilde{c})$ with probability at least $1 - \rho$, $G_0(\mu_\infty, \nu) \leq G_\delta(\mu_n, \nu) + \varepsilon(n, \rho, \zeta, K_\lambda) + \lambda\delta$. Also for $k > 1, d \geq 2, \zeta > 0, \delta > 0, \lambda = \delta^{-\frac{k-1}{k}}$ with probability at least $1 - \rho$, $G_0(\mu_\infty, \nu) \leq G_\delta(\mu_n, \nu) + \varepsilon(n, \rho, \zeta, K_\lambda) + \left(2 \cdot L^{\frac{k}{k-1}}(\widetilde{c}) + 1\right) \delta^{\frac{1}{k}}$.

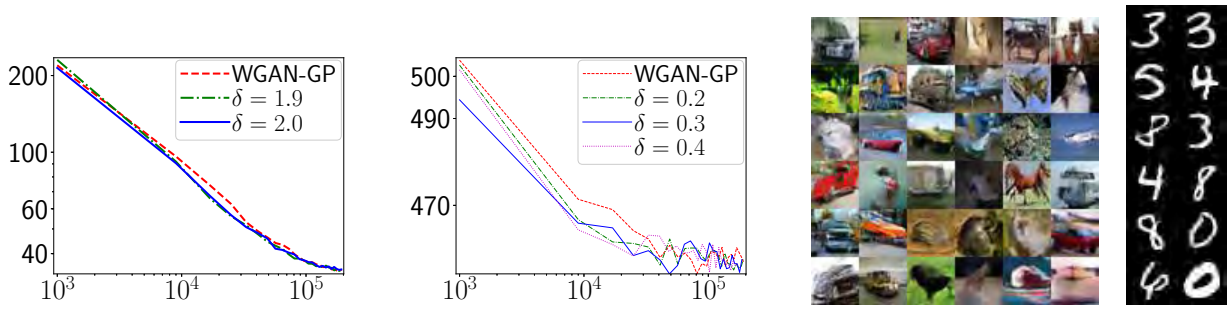
Remark 3 Theorem 4 also holds when μ_n and μ_∞ are switched. Hence, with probability at least $1 - 2\rho$, $G_0(\mu_\infty, \nu)$ resides in an interval centered at $G_\delta(\mu_n, \nu)$ with radius $\varepsilon(n, \rho, \zeta, K_\lambda) + \lambda\delta$ for $k = 1$ and radius $\varepsilon(n, \rho, \zeta, K_\lambda) + \left(2 \cdot L^{\frac{k}{k-1}}(\widetilde{c}) + 1\right) \delta^{\frac{1}{k}}$ for $k > 1$.

Remark 4 By optimizing the upper bound in Theorem 4, we were able to recover the term $\frac{1}{n^{1/d}}$ (curse of dimensionality) (Dudley 1969). For more details, see the appendix for Theorem 4.

4 EXPERIMENTS

OTR Wasserstein GAN This section provides experiment results evaluating OTR WGANs from Section 2. We trained on two dataset: MNIST (LeCun 1998) and CIFAR10 (Krizhevsky and Hinton 2009). For the WGAN implementation, we used the WGAN-GP code provided by (Gulrajani et al. 2017) and for Frechet Inception Distance (FID) calculation we used the code provided by (Heusel et al. 2017). For every fixed initial weights (seed), we trained our OTR WGAN with different values of δ. We performed training for 200000 generator iterations. For CIFAR10, $\delta \in \{0, 1.9, 2.0, 2.1\}$. For MNIST, $\delta \in \{0, 0.2, 0.3, 0.4\}$.

Representative training results are provided in Figures 1a, 1b in log-log scale. Sample images generated by OTR WGAN are provided in Figures 1c, 1d. Our experiments indicate that with an 'appropriate' choice of δ, OTR WGAN has a similar test loss performance to WGAN-GP. Also on the CIFAR10 dataset, they have similar Inception Score (Salimans et al. 2016) performance. Moreover, OTR WGAN has either the same or faster FID (Heusel et al. 2017) convergence rate than WGAN-GP. In addition, OTR WGAN trains faster than WGAN-GP because it skips training the Generator when the threshold criteria is not met. The 'appropriate' values for δ were found using cross-validation. This 'appropriate' value for δ should be slightly greater than $\min_\theta \mathscr{W}(\mu_n, \mu_\theta)$. For values of δ considerably larger than $\min_\theta \mathscr{W}(\mu_n, \mu_\theta)$, training of OTR WGAN is faster; however, the FID performance of OTR WGAN is worse than WGAN-GP. On the other hand for values of δ considerably less than $\min_\theta \mathscr{W}(\mu_n, \mu_\theta)$, the thresholding becomes ineffective and OTR WGAN behaves similar to WGAN-GP. In addition, trying many different initial points (seeds) indicate that OTR WGAN is more stable and has less volatility compared to WGAN-GP.



(a) (b) (c) (d)

Figure 1: FID versus generator iteration for comparison of OTR WGAN and WGAN-GP: (a) CIFAR10 (b) MNIST. In the legend, δ is the OTR WGAN parameter. Sample images generated by OTR WGAN: (c) with $\delta = 2.0$ when trained on CIFAR10 (d) with $\delta = 0.3$ when trained on MNIST.

Estimating the Optimal Transport Cost In this section, we present simulation results denoting the value of optimal transport regularization for estimating the Wasserstein distance between measures.

Let μ, ν be two probability measures defined on $\mathbb{R}^{20}$. The measure ν is constructed from 300 i.i.d samples of $\mathscr{N}(0, I_{20\times 20})$ where $I_{20\times 20}$ is the identity matrix. The measure μ is also constructed from 300 i.i.d. sampling of a random vector $X \in \mathbb{R}^{20}$ defined as follows. For each component X_i of X $(1 \le i \le 20)$, $X_i := \rho R_i + (1-\rho^2)^{\frac{1}{2}} T$ where $\{R_i\}_{i=1}^{20}, T$ are i.i.d. and $\mathscr{N}(0,1)$. In particular, $0 \le \rho \le 1$ specifies dependence of the X_i's.

For $n \in \mathbb{N}$, let μ_n be an empirical probability measure constructed from n i.i.d. samples from μ. For $\widetilde{c} = c = \|\cdot\|_2^2$, we compute the values of $G_{\delta_n}(\mu_n, \nu), G_0(\mu_n, \nu)$ where $\delta_n = \frac{1}{n^{0.45}}$. Then we compare them with the value of $G_0(\mu, \nu)$. To solve the optimal transport problems, the implementation from the Python Optimal Transport Library (Flamary, R'emi and Courty, Nicolas 2017) was used.

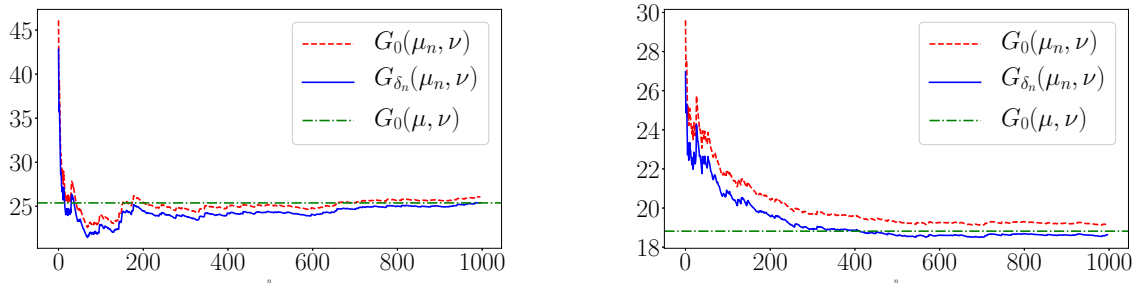
Our experiments indicate for large enough values of n, the empirical cost functions $G_0(\mu_n, \nu), G_{\delta_n}(\mu_n, \nu)$ often incur upward shifts relative to $G_0(\mu, \nu)$. This is illustrated in 2. Figure 2a and Figure 2b correspond to high and low dependence of the X_i's, respectively.

ACKNOWLEDGEMENTS

Support from NSF grants DMS-1720451 and DMS-1820942 as well as AFOSR MURI FA9550-20-1-0397 is gratefully acknowledged.

A PROOF OF THEOREM 1

The result follows directly from Sion's min-max theorem (see (Sion 1958)). First, the set $\mathscr{D}_\delta(\mu_0)$ is convex because, by duality, $D_c(\mu_0, \cdot)$ is convex (because since it is the supremum of linear functionals). Next,



(a) $\rho = 0.3$ (b) $\rho = 0.8$

Figure 2: Estimating the Optimal Transport Cost.

since the spaces involved are compact, the set $\mathscr{D}_\delta(\mu_0)$ is compact in the weak convergence topology, by Prohorov's theorem. Furthermore, it is immediate that the set $\mathscr{A}(\widetilde{c})$ is convex. Finally, the objective function is bilinear both in (μ,ν), on one hand, and (α,β) on the other. By definition of weak convergence, the functional is continuous in the weak convergence topology since the elements in $\mathscr{A}(\widetilde{c})$ are both continuous, and bounded and the spaces are compact.

B PROOF OF THEOREM 2

We have $G_\delta(\mu_0,\nu) = \min_{D_c(\mu,\mu_0)\leq\delta} \quad \min_{\pi\in\Pi_{X,Y}(\mu,\nu)} \quad \mathbb{E}_\pi \widetilde{c}(X,Y)$. Given a coupling $\pi\in\Pi_{X,Y}(\mu,\nu)$ we can always have a coupling between X and $W\sim\mu_0$ (by the gluing lemma, see (Villani 2003)). Therefore, we have that $G_\delta(\mu_0,\nu) = \min_\pi\{\int \widetilde{c}(x,y)\pi(dx,dy,dw) : \int c(x,w)\pi(dx,dw)\leq\delta,\ \pi_W=\mu_0,\ \pi_Y=\nu\}$. So,

$$-G_\delta(\mu_0,\nu) = \max_\pi \min_{\lambda\geq 0, h_1\in C(\mathscr{S}_X), h_2\in C(\mathscr{S}_Y)} \left[\int -\widetilde{c}(x,y)\pi(dx,dy,dw) + \int h_1(w)\mu_0(dw) - \right.$$
$$\left. \int h_1(w)\pi(dx,dy,dw) + \int h_2(y)\nu(dy) - \int h_2(y)\pi(dx,dy,dw) + \lambda\left(\delta - \int c(x,w)\pi(dx,dw)\right)\right].$$

Further, Sion's min-max Theorem (Sion 1958) is applicable because the value function is both linear in π and (h_1,h_2). In particular, it is concave in π and convex in (h_1,h_2). We then need to argue upper semicontinuity as function of π and lower semicontinuity as a function of (h_1,h_2). We choose the topology of uniform convergence over the compact sets $\mathscr{S}_X$ and $\mathscr{S}_Y$. Continuity then follows easily by Dominated Convergence. Now, to show upper semicontinuity as a function of π, we consider the space of probabilities under the weak convergence topology. It suffices to show that $\int \widetilde{c}(x,y)\pi(dx,dy,dw)$ and $\int c(x,w)\pi(dx,dw)$ are lower semicontinuous as a function of π, since the remaining terms involving π involve integrals of continuous functions over compact sets (hence continuous and bounded functions) and therefore those remaining terms are directly seen to be continuous by the definition of weak convergence (denoted by $\Rightarrow$). We need to show that if $\pi_n\Rightarrow\pi$ as $n\to\infty$, then $\liminf\int cd\pi_n \geq \int cd\pi$. By the Skorokhod representation, we may assume that there exists $Z_n = (X_n,Y_n,W_n)$ such that Z_n has distribution π_n and Z having distribution π, such that $Z_n\to Z$ almost surely as $n\to\infty$. Then, we have that $\liminf\int cd\pi_n = \liminf E(c(Z_n)) \geq \int E(\liminf c(Z_n)) \geq E(c(Z)) = \int c\, d\pi$, where the first inequality follows by Fatou's lemma and the second inequality follows because c is lower semicontinuous. A similar argument holds for $\widetilde{c}$. As a result,

$$-G_\delta(\mu_0,\nu) = \min_{\lambda\geq 0, h_1\in C(\mathscr{S}_X), h_2\in C(\mathscr{S}_Y)} \max_\pi \left[\int \left(-\widetilde{c}(x,y) - h_1(w) - h_2(y) - \lambda c(x,w)\right)\pi(dx,dy,dw)\right.$$
$$\left. +\lambda\delta + \int h_1(w)\mu_0(dw) + \int h_2(y)\nu(dy)\right].$$

The above expression implies that for all x,y,w we must have: $-\widetilde{c}(x,y)-h_1(w)-h_2(y)-\lambda c(x,w)\leq 0 \Rightarrow \sup_x[-\widetilde{c}(x,y)-\lambda c(x,w)]\leq h_1(w)+h_2(y)$. Therefore, the desired expression is obtained from the following.

$$-G_\delta(\mu_0,\nu)=\min_{\lambda\geq 0}\max_{\pi\in\Pi_{W,Y}(\mu_0,\nu)}\left\{\lambda\delta+\mathbb{E}_{\mu_0}h_1(W)+\mathbb{E}_\nu h_2(Y)\right\}$$

$$=\min_{\lambda\geq 0}\max_{\pi\in\Pi_{W,Y}(\mu_0,\nu)}\left\{\lambda\delta+\mathbb{E}_\pi\left[\sup_x\{-\widetilde{c}(x,Y)-\lambda c(x,W)\}\right]\right\}$$

C PROOF OF THEOREM 3

The differentiability of $h(\cdot)$ in λ is an immediate result of Corollary 4 of (Milgrom and Segal 2002). Define $u(\lambda,\mu_0,\nu)=\max_{\pi\in\Pi_{W,Y}(\mu_0,\nu)}\quad \mathbb{E}_\pi[h(W,Y,\lambda)]$. Therefore, $g(\lambda,\mu_0,\nu)=\delta\lambda+u(\lambda,\mu_0,\nu)$. In addition, define $f(\pi,\lambda)=\mathbb{E}_\pi[h(W,Y,\lambda)]$. We need to show $\frac{\partial}{\partial\lambda}u(\lambda,\mu_0,\nu)=\mathbb{E}_{\pi^*_\lambda}\left[\frac{\partial}{\partial\lambda}h(W,Y,\lambda)\right]$. For this statement to hold, according to Corollary 4 of (Milgrom and Segal 2002), a sufficient condition is as follows. The set $\Pi_{W,Y}(\mu_0,\nu)$ needs to be compact, $f(\pi,\lambda)$ needs to be continuous in π, and $\frac{\partial}{\partial\lambda}f(\pi,\lambda)$ needs to be continuous in (π,λ). In the remainder of this proof, we will show that this sufficient condition holds.

The set $\mathscr{S}_X\times\mathscr{S}_Y$ is compact. Therefore, Prohorov theorem implies under the weak convergence topology, $\Pi_{W,Y}(\mu_0,\nu)$ is a compact set.

On the other hand, the function $-\widetilde{c}(x,y)-\lambda c(x,w)$ is continuous in (x,w,y,λ) and the supremum $\sup_x\{-\widetilde{c}(x,y)-\lambda c(x,w)\}$ is attained due to the compactness of $\mathscr{S}_X,\mathscr{S}_Y$. Define $x^*(w,y,\lambda)$ to be the maximizing x, which will be unique (because $\widetilde{c}(\cdot,y)+\lambda c(\cdot,w)$ is strictly convex). Hence, $\sup_x\{-\widetilde{c}(x,y)-\lambda c(x,w)\}=-\widetilde{c}(x^*(w,y,\lambda),y)-\lambda c(x^*(w,y,\lambda),w)$. Then, Berge's maximum theorem (Aliprantis and Border 2006) implies $h(w,y,\lambda)$ is continuous in (w,y,λ) and $x^*(w,y,\lambda)$ is upper hemicontinuous in (w,y,λ). Moreover, since x^* is a single valued correspondence, it is continuous in (w,y,λ).

Since it is defined on a compact set and continuous, $h(\cdot,\cdot,\lambda)$ is bounded. Also, $f(\cdot,\lambda)$ is linear in π. Therefore, under the weak convergence topology, $f(\pi,\lambda)$ is continuous in π.

Moreover, $\frac{\partial}{\partial\lambda}f(\pi,\lambda)\overset{(a)}{=}\mathbb{E}_\pi\left[\frac{\partial}{\partial\lambda}h(W,Y,\lambda)\right]=\mathbb{E}_\pi[-c(x^*(W,Y,\lambda),W)]$. In this statement, (a) is an immediate result of the fact that $h(\cdot)$ is convex in λ and the monotone convergence theorem together with the fact that $h(\cdot)$ is differentiable in λ. Since c,x^* are continuous, the function $-c(x^*(W,Y,\lambda),W)$ is continuous in (W,Y,λ). For fixed λ, this function is bounded since it is defined on a compact set. Thus the bounded convergence theorem implies $\frac{\partial}{\partial\lambda}f(\pi,\lambda)$ is continuous in λ. In addition, $\frac{\partial}{\partial\lambda}f(\pi,\lambda)$ is continuous in π under the weak convergence topology. So, $\frac{\partial}{\partial\lambda}f(\pi,\lambda)$ is continuous in (π,λ).

D PROOF OF THEOREM 4 AND ADDITIONAL COMMENTS

Proof of Theorem 4 It can be shown (Villani 2003) that $f(X_1,\cdots,X_n):=\min_{\pi\in\Pi_{W,Y}(\mu_n,\nu)}\mathbb{E}_\pi\{-h(W,Y,\lambda)\}=\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\mu_n}\alpha(W)+\mathbb{E}_\nu\alpha^h_\lambda(Y)\}$ where $X_1,\cdots,X_n$ are the i.i.d samples associated with the empirical measure μ_n. $Lip(K_\lambda)$ denotes the set of all K_λ-Lipschitz functions $f(\cdot)$ defined on $\mathscr{S}_X$ such that $\min_{x\in\mathscr{S}_X}|f(x)|=0$. In addition, $\alpha^h_\lambda(y):=\sup_w\{-h(w,y,\lambda)-\alpha(w)\}$.

Proposition 1 For all $t>0$, $\mathbb{P}(f(X_1,\cdots,X_n)-\mathbb{E}f(X_1,\cdots,X_n)\geq t)\leq\exp\left(\frac{-2nt^2}{K_\lambda^2.diam^2(\mathscr{S}_X)}\right)$.

Using McDiarmid's inequality (Boucheron et al. 2013), to prove Proposition 1 it suffices to show that $f(\cdot)$ satisfies the bounded difference condition. Let $X_1,\ldots,X_n,X'_n$ be i.i.d samples from the measure μ_0. Let μ_n,μ'_n be the empirical measures associated with $X_1,\ldots,X_{n-1},X_n$ and $X_1,\ldots,X_{n-1},X'_n$ respectively.

$$|f(X_1,\ldots,X_n)-f(X_1,\ldots,X'_n)|=\left|\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\mu_n}\alpha(W)+\mathbb{E}_\nu\alpha^h_\lambda(Y)\}-\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\mu'_n}\alpha(W)+\mathbb{E}_\nu\alpha^h_\lambda(Y)\}\right|$$

$$\leq\left|\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\mu_n}\alpha(W)-\mathbb{E}_{\mu'_n}\alpha(W)\}\right|=\left|\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\frac{\alpha(X_n)-\alpha(X'_n)}{n}\right|\leq\frac{K_\lambda}{n}d(X_n,X'_n)\leq\frac{K_\lambda}{n}diam(\mathscr{S}_X)$$

Proposition 2 With probability at least $1-\rho$,

$$\min_{\pi\in\Pi_{W,Y}(\mu_n,\nu)}\mathbb{E}_\pi\{-h(W,Y,\lambda)\}-\min_{\pi\in\Pi_{W,Y}(\mu_0,\nu)}\mathbb{E}_\pi\{-h(W,Y,\lambda)\}\le\sqrt{\frac{\log(\frac{1}{\rho})}{2n}}+2R_n(Lip(K_\lambda))$$

where $R_n(\cdot)$ presents the Rademacher Complexity. The inequality also holds if μ_n,μ_0 are swapped.

Proof. $\min_{\pi\in\Pi_{W,Y}(\hat{\mu}_n,\nu)}\mathbb{E}_\pi\{-h(W,Y,\lambda)\}-\min_{\pi\in\Pi_{W,Y}(\mu_0,\nu)}\mathbb{E}_\pi\{-h(W,Y,\lambda)\}=$

$$\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\hat{\mu}_n}\alpha(W)+\mathbb{E}_\nu\alpha^h_\lambda(Y)\}-\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\mu_0}\alpha(W)+\mathbb{E}_\nu\alpha^h_\lambda(Y)\}$$

Using Proposition 1, with probability at least $1-\rho$, the above expression is less than or equal to

$$\sqrt{\frac{\log(\frac{1}{\rho})}{2n}}+\mathbb{E}\left[\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\hat{\mu}_n}\alpha(W)+\mathbb{E}_\nu\alpha^h_\lambda(Y)\}-\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\mu_0}\alpha(W)+\mathbb{E}_\nu\alpha^h_\lambda(Y)\}\right]$$
$$\le\sqrt{\frac{\log(\frac{1}{\rho})}{2n}}+\mathbb{E}\left[\sup_{\alpha(\cdot)\in Lip(K_\lambda)}\{\mathbb{E}_{\hat{\mu}_n}\alpha(W)-\mathbb{E}_{\mu_0}\alpha(W)\}\right]\overset{(a)}{\le}\sqrt{\frac{\log(\frac{1}{\rho})}{2n}}+2R_n(Lip(K_\lambda))$$

where (a) is based on the first inequality in Section 5 of (Luxburg and Bousquet 2004). The proof of the other inequality is similar. □

Define $q_k=\begin{cases}\lambda\delta, & k=1\\ \left(2\cdot L^{\frac{k}{k-1}}(\widetilde{c})+1\right)\delta^{\frac{1}{k}}, & k>1\end{cases}$. Now for the event of interest we have

$$\begin{aligned}
&\{G_0(\mu_0,\nu)\le G_\delta(\mu_n,\nu)+\varepsilon(n,\delta,\zeta,K_\lambda)+q_k\}\\
&=\left\{q_k\ge G_0(\mu_0,\nu)+\min_{\lambda\ge0}\left\{\delta\lambda+\max_{\pi\in\Pi_{W,Y}(\mu_n,\nu)}\mathbb{E}_\pi h(W,Y,\lambda)\right\}-\varepsilon(n,\delta,\zeta,K_\lambda)\right\}\\
&=\Bigg\{\exists\lambda\ge0: q_k\ge\delta\lambda+\underbrace{\max_{\pi\in\Pi_{W,Y}(\mu_n,\nu)}\mathbb{E}_\pi h(W,Y,\lambda)-\max_{\pi\in\Pi_{X,Y}(\mu_0,\nu)}\mathbb{E}_\pi h(X,Y,\lambda)-\varepsilon(n,\delta,\zeta,K_\lambda)}_{(I)}\\
&\quad+\underbrace{\min_{\pi\in\Pi_{X,Y}(\mu_0,\nu)}\mathbb{E}_\pi\widetilde{c}(X,Y)+\max_{\pi\in\Pi_{X,Y}(\mu_0,\nu)}\mathbb{E}_\pi h(X,Y,\lambda)}_{(II)}\Bigg\}.
\end{aligned}$$

Below, we show that the above event occurs with probability at least $1-\rho$.

Lemma 1 Let $(\mathscr{S},d)$ present a compact metric space. Also let $f:\mathscr{S}\to\mathbb{R}$ be an L-Lipschitz function (for $L>0$) (i.e. for $x,y\in\mathscr{S}$, $|f(x)-f(y)|\le L\cdot d(x,y)$). For $k\ge1,\lambda>0$, define $y_x:=\arg\max_{y\in\mathscr{S}}\{f(y)-\lambda\cdot d^k(x,y)\}$. Then for $k=1$ and $\lambda>L$, $y_x=x$. Also for $k>1$, $d(y_x,x)\le(L/\lambda)^{1/(k-1)}$.

Lemma 1 is an immediate result of the following.

$$f(x)=f(x)-\lambda\cdot d^k(x,x)\le f(y_x)-\lambda\cdot d^k(x,y_x)\Leftrightarrow\lambda\cdot d^k(x,y_x)\le f(y_x)-f(x)\le L\cdot d(x,y_x)$$
$$\Leftrightarrow(\lambda d^{k-1}(x,y_x)-L)d(x,y_x)\le0$$

For (I), using Proposition 2 with probability at least $1-\rho$:

$$\max_{\pi\in\Pi_{X,Y}(\mu_n,\nu)} \mathbb{E}_\pi h(X,Y,\lambda) - \max_{\pi\in\Pi_{X,Y}(\mu_0,\nu)} \mathbb{E}_\pi h(X,Y,\lambda) - \varepsilon(n,\delta,\zeta,K_\lambda)$$
$$\leq \sqrt{\frac{\log(\frac{1}{\rho})}{2n}} + R_n(Lip(K_\lambda)) - \varepsilon(n,\delta,\zeta,K_\lambda) \overset{(a)}{\leq} 0$$

where (a) is due to Theorem 18 of (Luxburg and Bousquet 2004). For (II):

$$\min_{\pi\in\Pi_{X,Y}(\mu_0,\nu)} \mathbb{E}_\pi \widetilde{c}(X,Y) + \max_{\pi\in\Pi_{X,Y}(\mu_0,\nu)} \mathbb{E}_\pi h(X,Y,\lambda) = \max_{\pi\in\Pi_{X,Y}(\mu_0,\nu)} \mathbb{E}_\pi h(X,Y,\lambda) - \max_{\pi\in\Pi_{X,Y}(\mu_0,\nu)} \mathbb{E}_\pi\{-\widetilde{c}(X,Y)\}$$
$$\leq \max_{\pi\in\Pi_{X,Y}(\mu_0,\nu)} \mathbb{E}_\pi\{h(X,Y,\lambda)+\widetilde{c}(X,Y)\}.$$

For $k=1$, Lemma 1 indicates $h(x,y,\lambda)+\widetilde{c}(x,y)=0$ for $\lambda > L(\widetilde{c})$ and all (x,y). This concludes the proof for $k=1$. For $k>1$, Lemma 1 shows that for all (x,y), $h(x,y,\lambda)+\widetilde{c}(x,y) \leq L(\widetilde{c})\,(L(\widetilde{c})/\lambda)^{\frac{1}{k-1}} + \lambda\,(L(\widetilde{c})/\lambda)^{\frac{k}{k-1}} = 2L^{\frac{k}{k-1}}(\widetilde{c})/\lambda^{\frac{1}{k-1}}$. Now setting $\lambda = (\frac{1}{\delta})^{\frac{k-1}{k}}$, we get $\delta\lambda + (II) \leq (2\cdot L^{\frac{k}{k-1}}(\widetilde{c})+1)\delta^{\frac{1}{k}} = q_k$.

Additional Comments From Theorem 18 of (Luxburg and Bousquet 2004), for connected and centered sets $\mathscr{S}_X$, with the following (tighter) definition for $\varepsilon(n,\rho,\zeta,K_\lambda)$, Theorem 4 still holds.

$$\sqrt{\frac{\log(\frac{1}{\rho})}{2n}} + 4\zeta K_\lambda + \frac{8\sqrt{2}K_\lambda}{\sqrt{n}} \int_{\zeta/4}^{2diam(\mathscr{S}_X)} \sqrt{\mathscr{N}(\mathscr{S}_X,d_X,\xi/2)\log 2 + \log\left(2\left\lceil\frac{2diam(\mathscr{S}_X)}{\xi}\right\rceil+1\right)}d\xi$$

In particular when $\mathscr{S}_X=[0,1]^d$ and d_X is the Euclidean metric, $\mathscr{N}(\mathscr{S}_X,d_X,\xi) \leq \frac{H_d}{\xi^d}$ for $\xi\leq 1$ and some $H_d>0$. So, minimizing ζ for $d>2$ and sufficiently large values of n results in

$$\varepsilon(n,\rho,\zeta,K_\lambda) \leq \sqrt{\frac{\log(\frac{1}{\rho})}{n}} + \left(\frac{32K_\lambda d}{d-2}\right)\left(\frac{H_d\cdot\log 2}{2n}\right)^{1/d} + \frac{(8\sqrt{2})(8+2\sqrt{H_d\cdot\log 2})diam(\mathscr{S}_X)K_\lambda}{\sqrt{n}}.$$

The $n^{-1/d}$ factor (curse of dimensionality) aligns with (Dudley 1969; Weed and Bach 2019).

REFERENCES

Abadeh, S. S., P. M. M. Esfahani, and D. Kuhn. 2015. "Distributionally robust logistic regression". In *Advances in Neural Information Processing Systems*, 1576–1584.

Aliprantis, C. D., and K. C. Border. 2006. *Infinite Dimensional Analysis*. Third ed. Springer.

Arjovsky, M., S. Chintala, and L. Bottou. 2017. "Wasserstein generative adversarial networks". In *International Conference on Machine Learning*, 214–223.

Bertsekas, D. P. 2015. *Convex optimization algorithms*. Athena Scientific.

Blanchet, J., Y. Kang, and K. Murthy. 2019. "Robust Wasserstein profile inference and applications to machine learning". *Journal of Applied Probability* 56(3):830–857.

Blanchet, J., and K. Murthy. 2019. "Quantifying distributional model risk via optimal transport". *Mathematics of Operations Research*.

Boucheron, S., G. Lugosi, and P. Massart. 2013. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press.

Cuturi, M. 2013. "Sinkhorn distances: Lightspeed computation of optimal transport". In *Advances in Neural Information Processing Systems*, 2292–2300.

Dudley, R. 1969. "The speed of mean Glivenko-Cantelli convergence". *The Annals of Mathematical Statistics* 40(1):40–50.

Esfahani, P. M., and D. Kuhn. 2018. "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations". *Mathematical Programming* 171(1-2):115–166.

Flamary, R'emi and Courty, Nicolas 2017. "POT Python Optimal Transport library".

Forrow, A., J.-C. Hütter, M. Nitzan, P. Rigollet, G. Schiebinger, and J. Weed. 2019. "Statistical optimal transport via factored couplings". In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2454–2465.

Gao, R., X. Chen, and A. J. Kleywegt. 2017. "Wasserstein distributional robustness and regularization in Statistical learning". *arXiv preprint arXiv:1712.06050*.

Genevay, A., M. Cuturi, G. Peyré, and F. Bach. 2016. "Stochastic optimization for large-scale optimal transport". In *Advances in Neural Information Processing Systems*, 3440–3448.

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. 2017. "Improved training of Wasserstein GANs". In *Advances in Neural Information Processing Systems*, 5767–5777.

Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. 2017. "GANs trained by a two time-scale update rule converge to a local Nash equilibrium". In *Advances in Neural Information Processing Systems*, 6626–6637.

Kolouri, S., S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. 2017. "Optimal mass transport: Signal processing and machine-learning applications". *IEEE Signal Processing Magazine* 34(4):43–59.

Krizhevsky, A., and G. Hinton. 2009. "Learning multiple layers of features from tiny images". Technical report.

Kusner, M., Y. Sun, N. Kolkin, and K. Weinberger. 2015. "From word embeddings to document distances". In *International Conference on Machine Learning*, 957–966.

LeCun, Y. 1998. "The MNIST database of handwritten digits". *http://yann. lecun. com/exdb/mnist/*.

Luxburg, U. v., and O. Bousquet. 2004. "Distance-based classification with Lipschitz functions". *Journal of Machine Learning Research* 5(Jun):669–695.

Milgrom, P., and I. Segal. 2002. "Envelope theorems for arbitrary choice sets". *Econometrica* 70(2):583–601.

Nesterov, Y. 2005. "Smooth minimization of non-smooth functions". *Mathematical Programming* 103(1):127–152.

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. 2016. "Improved techniques for training GANs". In *Advances in Neural Information Processing Systems*, 2234–2242.

Sanjabi, M., J. Ba, M. Razaviyayn, and J. D. Lee. 2018. "On the convergence and robustness of training GANs with regularized optimal transport". In *Advances in Neural Information Processing Systems*, 7091–7101.

Sion, M. 1958. "On general minimax theorems.". *Pacific Journal of Mathematics* 8(1):171–176.

Villani, C. 2003. *Topics in Optimal Transportation*. Number 58. American Mathematical Soc.

Weed, J., and F. Bach. 2019. "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance". *Bernoulli* 25(4A):2620–2648.

AUTHOR BIOGRAPHIES

Saied Mahdian is a PhD candidate in Operations Research at Stanford University. His email is smehdian@stanford.edu.

Jose H. Blanchet is a Professor at Stanford University. His research interests lie in stochastic modeling, simulation and stochastic optimization. His email is jblanche@stanford.edu.

Peter W. Glynn is the Thomas Ford Professor at Stanford University. His research interests lie in simulation, computational probability, statistical inference. His email is glynn@stanford.edu.